A Mixed Methods Study on Formative Assessment of EFL Writing of College Students in Japan: Focusing on the Effectiveness of Schematized Teacher Feedback on Coherence

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Education in the Graduate School of Education

Kana Matsumura

Waseda University

2023

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deep and sincere gratitude to my research supervisor, Dr. Yasuyo Sawaki for giving me the opportunity to conduct research and for providing invaluable guidance throughout this research. It has been a great privilege and honor to work and study under her guidance. I am extremely grateful for what she has offered me and her patience during the discussion I had with her about the research work and thesis preparation. I am also most grateful to Professor Emeritus Dr. Michiko Nakano. She is the benefactor who gave me the opportunity to return to the academic world after a long period of hiatus in my research life. Her precise and enlightening advice on research has always given me great insight.

I would like to express my sincere appreciation to Sawaki Seminar, Nakano Seminar and all my research colleagues who have collaborated with me over the years. Their presence has always given me the courage and motivation to continue my research.

Finally, I would like to express my sincere gratitude to my family. My two daughters, Miki and Yukina, have always been right by my side and have always supported me when I felt discouraged. In closing, I would like to thank my parents, my husband's parents, and my dearest husband, Koki, who are in heaven, for always watching over me. I have always felt their presence in my heart.

I would not be where I am today if I had lacked any of the people I have met in my life.

Chapter 1

Introduction

1.1 Background and Research Purpose

Writing argumentative paragraphs in English requires English-as-a-foreign-language (EFL) learners to state their opinions by following certain discourse patterns that may differ from those of their first language (e.g., Connor, 1996; Kaplan, 1966; Oi, 1986, 1999, 2005). Montgomery and Baker (2007) remark that the goal of L2 writing is to teach/learn the writing convention of a particular culture as well as the grammatical forms of the target language. EFL learners are expected to learn not only to produce grammatically correct sentences in L2 writing but to manipulate English language organization to present a strong argument that appeals to English language readers. It is undeniably important to produce a passage consisting of a group of sentences that are connected coherently to communicate a clear and persuasive message. The ability to present the writer's opinion with relevant supporting detail to develop an argument systematically in written form is required in the academic domain (Council of Europe, 2001). This ability to construct ideas in well-organized structure following the English language format is a basic skill that is also needed in the general domain. Learning how to write paragraphs, the basic unit of coherent sentences in English, is an important skill for EFL learners that must be acquired at the introductory level so that they can write reports and papers in the academic domain in the years ahead. This study empirically investigated the teaching and learning of coherent English paragraph writing at a basic level in universities in Japan.

In standardized tests such as the Test of English as Foreign Language internet Based Test (TOEFL iBT) administered by English Testing Service (ETS) and the International English Language Testing System (IELTS) tests administered by British

Council and the International Development Program (IDP) Education, which are designed to measure the English proficiency required to survive in a university classroom or as a citizen in an English-speaking country, independent writing tasks require the test takers to present their opinion or to develop an argument, and their products must display unity, progression, and coherence to earn higher scores. Consequently, EFL learners in higher education need some practice to acquire this language skill to produce a well-organized paragraph, namely *coherence*.

Coherence is defined in this study as sentences that are arranged in proper order, seamlessly connected, and in which ideas are presented one after another without stagnation both within and between units. Coherence is considered a significant quality of effective writing (Bamberg, 1984; Richards, 1990; Wilkinson, 1990). At the same time, coherence is often considered difficult to learn and teach in the classroom settings. Cerniglia *et al*. (1990) state that, despite most teachers' recognition of coherence as "an essential element of good writing," it is still difficult to teach, and "students still do not know how to write coherently" (p. 229). They also referred, in particular, to English-as-a-second-language (ESL) students' difficulties in writing coherent English texts. Along the lines of this, Lee (2002) remarks concerning coherence:

> In the writing classroom, ESL teachers tend to refer to "coherence" in abstract terms without making a systematic attempt to explain and to teach it, saying, for instance, *your writing is not coherent; your writing lacks unity; the ideas don't hang together; the ideas are disorganized* (p. 137).

As stated above, it is not always easy for teachers in the classroom to evaluate their students' writing to articulate problems with coherence. This naturally leads to students' difficulties in comprehending why certain writing earns a good/poor evaluation for coherence of discourse. This, however, should not be ascribed to contrastive rhetorical issues strictly related to their first language. Rather, it can be ascribed to difficulties in defining the criteria of coherence for evaluation, which may come from "vague descriptions that might lie in the rather vague nature of coherence itself" (Knoch, 2007a, p. 109). Moreover, Bamberg (1984) remarks that coherence can have multiple meanings, so the meaning when one writer refers to "coherence" may be different from the way another writer refers to it. As she explains further, coherence has dual perspectives: local and global coherence. In English composition classes in particular, local transitions or connections between sentences are often focused on and referred to as coherence, which Halliday and Hasan (1976) and van Dijk (1980) call cohesion. In contrast, global coherence refers to the overall text structure or organization at the discourse level. Furthermore, to generate global coherence, a semantic structure in which all propositions are linked together must be built according to a certain conventional framework (Bamberg, 1984; Shank & Abelson, 1977; van Dijk, 1980). When discussing coherence in writing, these rhetorical and ideational aspects must be considered.

The vagueness of coherence is also reflected in the descriptors in the criteria of large-scale language tests' rating scales (Matsumura & Takagi, 2022). Because coherence is often evaluated in terms of degrees of quality rather than simple correctness or incorrectness, they are described with adjectives or adverbs such as *well developed*, *poorly organized*, *generally clear*, *formulaic*, or *succinct*. These wordings are also used for feedback, which may possibly make recipients puzzled, specifically when they are used as formative assessment. The learners are not sure how to solve their problems in coherence when it comes to revising them in process writing activities in the classroom. Accordingly, there is a need to develop formative feedback provided to students in formative assessments that will help them address coherence issues during the revision work of process writing activities in the EFL classroom.

In an attempt to respond to the problems associated with teaching and learning ideational and rhetorical coherence, this study explores how to teach and learn coherence effectively in introductory English writing class in college through classroom-based formative language assessment. To address this issue, the present study attempts to generate and use diagnostic writing feedback with a schematized tree diagram developed by a web-based annotation tool. This study focuses on argumentative, stand-alone paragraph writing, which requires a certain framework and emphasizes logical coherence. The study utilizes an annotation tool that allows tagging sentences as text units as part of the formative assessment procedure of coherence and automatically generates the results as a hierarchical tree diagram. This approach was designed based on the hypothesis that it is possible to evaluate both local coherence between adjacent sentences through tagging and global coherence through a graphical representation that offers an overall view of the paragraph structure. This form of formative feedback is expected to facilitate coherence teaching and learning better than conventional text-based feedback that highlights problem areas with underlines and offers comments in the text.

1.2 Terminology Related to Coherence

As described in the previous section, there is some ambiguity in the use of terms related to coherence, so this section provides definitions of the following terms used in this study in terms of written text: *organization*, *cohesion*, *coherence*, and *coherence anomaly/break*. In addition, the definition of *argumentative writing* in the study is also presented at the end.

The term *organization* is used in various fields on a daily basis, not only in linguistics. As the general meaning of "organization" refers to the successful functioning of a single unit of individuals as a whole, it indicates a state of overall harmony (Pugh et al., 1969). In a language-related use, it generally refers to the entire structure of a text in a broad sense, including coherence, which is the central concept of this study. In terms of writing assessment, the term is widely used in the writing evaluation scale criterion as a term for the cohesion of sentences/how multiple sentences are organized, and unlike coherence and cohesion, it is often not explicitly defined separately. The characteristics of organization are described in rating scales employed in some previous studies as follows: Organized paragraph is clear and easy to follow (White, 1994), and well-organized text presents logical paragraphing (Diagnostic English Language Needs Assessment [DELNA]; Doe, 2014). Cohen (1994) remarks that organization in writing is evaluated in terms of a sense of patterns for the development of ideas. In a study of diagnostic assessment, Kim (2011) provides a good summary of the concept of organization in terms of writing assessment as organizational effectiveness:

> Organizational effectiveness assesses the way in which a writer organizes and develops his or her ideas. A writer who is competent in this area generally demonstrates the ability to construct and develop a paragraph effectively and to connect textual elements well within and between paragraphs using appropriate cohesive and transitional devices (p. 518).

To sum up, *organization* can be defined as how ideas are presented, referring to the larger parts of a piece of writing such as a paragraph or an essay so that it helps readers follow and understand the information presented. In that sense, coherence and cohesion are defined as components of text organization. However, the term "organization" in a rating scale is sometimes used synonymously with coherence, which refers to the narrower sense of organization. As such, the use of the term, organization, which sums up the various elements, is avoided in the present study to prevent confusion, unless used in the description of previous studies.

Cohesion[1] is a structural connection between sentences and their components that is explicitly illustrated with formal markers (van Dijk, 1977). Halliday and Hassan (1976) define the term as "part of the text-forming component in the linguistic system.

[1] *Cohesion* refers to the structural connection of the constituent elements between sentences with two types of connections: grammatical and lexical methods. The grammatical methods include elements such as co-reference, substitution, ellipsis, and connective relation, and the lexical ones include repetition, synonymy, hyponymy, and metonymy (Halliday & Hassan, 1976).

It is the means whereby elements that are structurally unrelated to one another are linked together, through the dependence of one on the other for its interpretation" (p. 27). In this dissertation, cohesion is regarded as a significant component of coherence.

Coherence in written text is a type of coherence called propositional coherence, which is based on the organization or the propositional content of discourse (Lautamatti, 1990), while the one in spoken discourse is called interactional coherence. Here, coherence is defined as "semantic property of discourses, based on the interpretation of other sentences" (van Dijk, 1977, p. 93). While cohesion is explicitly illustrated with formal markers, coherence shows more implicitly how the propositions are connected through sentences based on the knowledge of writers and readers with relations such as cause, recession, and motivation, among others.

A *coherence anomaly/break* is defined as "what happens when the reader loses the thread of the argument while in the process of reading a text attentively" (Wikborg, 1990, p. 133). Because it is a result and a phenom, it certainly presupposes a variety of causal factors.

Lastly, in this dissertation "argumentative writing" is defined as writing that expresses one's opinion and states at least one piece of evidence to support it, and with the purpose of convincing the reader, who may or may not have a dissenting opinion. The term is based on Crusius and Channell's (2004) categorization of arguments with Toulmin's point of view.

In this study, *cohesion*, which expresses an explicit structural connection, and *coherence*, which implicitly expresses a propositional transfer from the connection of text units, are treated as ideational and rhetorical coherence. In contrast, the term *organization*, which includes a variety of factors and has a wide range of definitions, is only used to describe previous studies, specifically in the rating scale of the ESL Composition Profile (hereafter ESL CP) described by Jacobs et al. (1981), on which previous quantitative studies rely, and in the descriptors of other previous studies.

1.3 Issues in Formative Assessment in EFL Writing Classes

This study focuses on formative language assessment and feedback in the classroom context to investigate the effectiveness of schematized feedback. This section briefly summarizes the fundamental issues of classroom formative assessment related to teaching and learning coherence in EFL writing.

1.3.1 What Is Required for Formative Feedback in Classroom-Based Assessment?

Formative assessment is also termed "classroom evaluation," "curriculum-based assessment," "feedback," and "formative evaluation," among others (Black & Wiliam, 1998, p. 53). This assessment is "to be interpreted as encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (Black & Wiliam, 1998, pp. 7-8). Formative assessment can also be argued from an alternative comparison between the actual performance and the reference level of a system parameter. In other words, when the actual level of the target examinee is compared with the reference level, there is a gap between the two. The assessment is considered formative when information is obtained to correct the gap, that is, to apply the information to proceed to the next step (Black & Wiliams, 1998; Ramaprasad, 1983; Sadler, 1989). This presupposes the fact that evaluation of the actual performance can be compared by using a reference standard, which would give diagnostic information on how far from a standard a performance was, or what should be done to achieve a reference standard. Actually, the activity of formative assessment consists of the following two sequential actions: the recognition by the learner of the gap between their current ability and the ability to which they are aiming, and the closing of that gap to achieve that desired goal. To promote this sequence of actions, it is often effective to have appropriate feedback from a rater or instructor.

Formative assessment of EFL writing includes the options of self-assessment and peer assessment, where the learner plays the role of the agent of assessment, and Oi

(2021) described the effectiveness of this assessment as part of English writing instruction. However, in the present study, teacher assessment and feedback were used due to the nature of the assessment implemented, which required text tagging based on specialized knowledge, and because the purpose of the task was to ensure efficient revision writing by students. Particularly in the rewriting task by students, in the case of performance tests such as speaking and writing, more detailed assessment is desired to bridge the gap between the students' performance and a reference standard because the rating is not based on a simple right/wrong score.

The main body of learning is the learner, but a teacher's feedback acts as a scaffolding to promote the learners' awareness of the current state of their own skills. In classroom assessment, the individual formative feedback provided by the teacher helps the learners recognize how close or far their skills are from a goal or a reference level. Therefore, first and foremost, teachers need to decide on a reference standard that meets their students' goal and descriptors they can use for classroom assessment. They design a formative assessment and task(s) that can evaluate their students' targeted language skills to generate sufficient feedback. In this sense, feedback is like a nutrient that teachers extract from the source data, sort, and share with students in an optimal way so they understand the information obtained from the formative language assessment. Formative feedback as a part of formative assessment is a catalyst that activates students' awareness and understanding of the targeted language use in the classroom context.

Feedback, needless to say, is essential to a student's writing development (Biber, Nekrasova, & Horn, 2011; Ferris, 2003; Hyland, 2003; Montgomery & Baker, 2007). In the process of formative assessment, various formative decisions are made, and providing feedback is one of them. For example, teachers have to make formative decisions at the right time and in the correct manner to provide appropriate feedback as it is directly related to development or improvement of teaching and learning. Formative decisions are decisions that lead to activities that are intended to improve instruction and learning often made on the basis of classroom-based assessment (Bachman & Damböck, 2017), and they are "intended to help students guide their own subsequent learning" as well as improvement of teachers' instructions (Bachman & Palmer, 2010, p. 197). Therefore, in the classroom setting, formative decisions are made by both teachers and their students at various times and situations. Based on the decisions, they interactively proceed with repeated exchanges of information between both sides, and they have the best effect when they interact harmoniously throughout the entire process.

Feedback is one of the significant constituents of assessment that helps promote harmonious interaction. Diagnostic feedback, which provides learners with detailed useful information, is expected to improve their comprehension of the assessment results to be utilized for the next step. In Lee and Sawaki's (2009) conceptual overview of the cognitive diagnosis approach (CDA)[2], they refer to a crucial feature of diagnosis from language assessment viewpoint as "identifying strengths and weaknesses of individual learners in the targeted area of learning and instruction" (p. 172) by suggesting the necessity of describing more fine-grained information about learners. They also point out the significance of detailed feedback to show the state of the learners' skills so that their teachers can take appropriate actions to improve their skills in the targeted language. More issues on diagnostic assessment are discussed later in the literature review in Chapter 2.

1.3.2 Issues in Writing Feedback in Terms of Assessment of Coherence

Formative decisions, in the writing classroom specifically, are often embedded in the process approach to writing, for which the development of students' writing skills

[2] Lee and Sawaki (2009) explain the CDA based on the previous studies on which they are cognitively grounded, namely diagnostic assessment procedures. These procedures usually follow four steps: definition of attributes, Q-matrix construction, data analysis, and score reporting/diagnostic feedback.

can be measured not only within a single draft, but also across drafts. In English writing classes at the college level in Japan, a process approach accompanied by the teacher's feedback seems to be one of the most commonly used instruction methods (Oi et al., 2016). Considering that a series of activities in the classroom is part of formative assessment, detailed feedback is not required every time. However, in major tasks such as writing drafts, it is necessary to provide students with comprehensive and diagnostic feedback so they can revise their drafts. Teachers, as test developers and decision-makers in the classroom, hold accountability for the uses of a particular assessment (Bachman & Palmer, 2010). This means that, if teachers evaluate their students' writing as *not being coherent* or *lacking unity*, they should be able to justify the reason(s) for such an evaluation. However, pinpointing the problems of coherence is not as easy as pointing out grammatical errors, inappropriate use of vocabulary, or mechanical errors. The reasons of this difficulty may be twofold: One is that teachers themselves have difficulties in detecting coherence anomalies consistently (Cumming, 1990). The second is that it is not easy to explain the problems to their students (Cerniglia et al., 1990; Lee, 2002).

The first reason for this difficulty to detect problems consistently in coherence may be ascribed to the considerably abstract definition of the organization criterion. Various viewpoints that the organization covers may result in dependence on the raters' subjective judgment of what they value (Knoch, 2007a; Todd et al., 2004). Unlike explicit errors in language use, the rater is required to have some sort of subjective ability to be sensitive to the presence of anomalies as a reader, such as a sense of strain or discomfort that stagnates the flow of the sentence. Moreover, the level of tolerance to coherence problems could vary depending on the values of the rater when the rating scale is described in terms of the degree of quality. It is no surprise that Knoch (2007a), who did not want the possibility of fluctuations in the evaluation criteria by such raters, decided to limit the evaluation of coherence to a centralized topical structure analysis (TSA) so that subjective judgment is minimized. Furthermore, it is not always easy for raters to identify a clear basis for an anomaly or the location of its occurrence because it involves multiple sentences and a larger context. Hence, there is difficulty in successfully expressing and communicating the identified anomaly.

The second reason for not communicating an anomaly well lies in the difficulty teachers have in clearly explaining the coherence problems to their students. This may be partly due to the feedback form or modality. In the writing class, the written mode of feedback includes several forms such as marginal comments, end comments, editing codes, and circles/underlines (Biber et al., 2011). These forms may be effective for corrective feedback on language use or mechanics, but they might not be practical enough to comment on organization across sentences from a micro- and macro-structural perspectives. The more detailed the feedback is meant to be with marginal comments next to the text, the more cumbersome the feedback tends to become. Nevertheless, if the feedback is summarized into end comments, it would probably be much less concrete. An oral mode of feedback might help teachers explain the problem across the text, but teachers still need to have some specific information suggesting the location of the problems, and students would want written feedback that they could review after the feedback session. These issues suggest the need to describe coherence anomalies or breaks identified to the learners.

In response to the above call for clear description of writing coherence, this dissertation attempts to use tree diagrams created with a computational tool as part of formative feedback. The Tool for Interactive Argument Annotation (TIARA) is a newly developed annotation tool by a research group from the Tokyo Institute of Technology (Putra et al., 2020). Chapter 3 provides detail on how the tool is used for assessment.

1.4 Theoretical Background and Research Design of the Present Study

To conclude the theoretical background of an investigation into formative language assessment, it is necessary to discuss the framework and the research approach. This study employs the assessment use argument (AUA) (Bachman &

Damböck, 2017; Bachman & Palmer, 2010) as the basis for justifying the use of language assessment. The AUA is a prominent argument-based approach to validation (Chapelle & Voss, 2014). To implement the validity argument in the AUA framework in practice, this study applies a mixed methods approach, where the investigator gathers both quantitative and qualitative data, and draws interpretations by integrating the two to understand research problems (Creswell, 2015). Fundamentally, a communicative language assessment has a high affinity to a mixed methods approach, as noted by Moeller (2016), the first editor of *Second Language Assessment and Mixed Methods Research*. Moeller (2016) remarks in the chapter of the book on the significance of a mixed methods approach in terms of a language assessment in the classroom that instruction and testing are often separable. Furthermore, implicit assessment continuously takes place in the classroom settings, and the information obtained through this type of assessment is subject to a mixed methods interpretation. She says, “The numerous variables and complexity in assessing authentic task-based communication at the classroom level in addition to the challenges such as reliability, content validity and authenticity (Bachman & Palmer, 2010; Norris, 2009; Wigglesworth, 2008) underscore that one research method cannot fully capture the complexity of language skills” (Moeller, 2016, p. 8). The mixed methods approach, which allows triangulation perspective, alongside the AUA framework, which includes multiple steps to validate the assessment, would lead to more theoretically robust conclusions and better accountability for the information derived from them than single research method. Table 1.1 shows the overview of the current research design. The left column shows the research constituents, and the right column shows the specifics of each component.

Table 1.1

Overview of the Current Research Design

Constituents	Detail
Context	Formative assessment in a classroom setting
Language use domain	Producing an organized and coherent paragraph in EFL argumentative writing
Purpose	To investigate the effectiveness of teacher's schematized feedback to teach and learn coherence in EFL paragraph writing
Means	A tool with which teachers can provide diagnostic feedback with graphical presentation
Instruments	(a) Analytic rating scale (b) Discourse annotation accompanied by annotated diagrams with an annotation tool (c) Questionnaires to students
Variables	(a) Analytic rating scores (b) Coherence anomalies/breaks (c) Responses obtained from questionnaires
Framework	The AUA from a mixed methods approach

1.4.1 Assessment Use Argument as the Validity Framework

The assessment of this study is approached based on the AUA proposed by Bachman and Palmer (2010). This framework provides specific guidelines for implementing the validity and usefulness verification in language testing with specific cases depending on the assessment context. Moreover, the latest book written by Bachman and Damböck (2017), *Language Assessment for Classroom Teachers*, which is also based on the AUA, supports the current research goal in the authentic classroom settings. In addition to the AUA, the Toulmin model of argument (Toulmin, 2003), which is the framework of practical reasoning that underlies it, is discussed here. The annotation scheme used in discourse analysis along with instruction of argumentative writing in the classroom are also based on this argument model. Before delving into the validity framework of the current research, a brief overview of the development of validity in language testing is provided by focusing on the perspectives of argument-based approach.

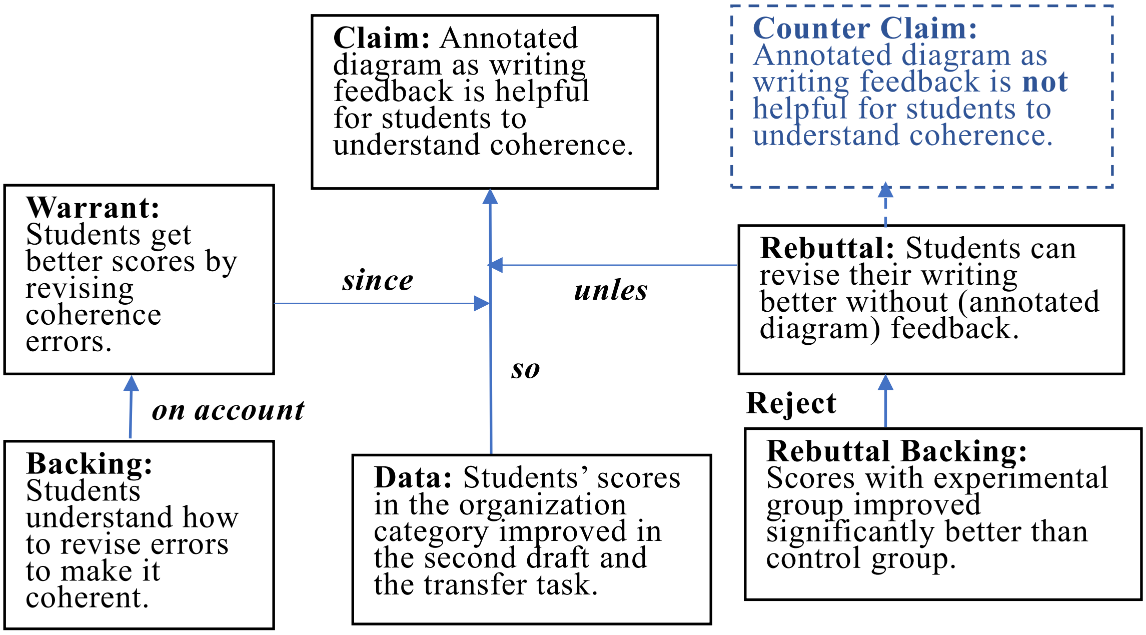
The argument-based approach has become mainstream in the language testing area, and many researchers have used this approach (e.g., Bachman & Palmer, 2010; Chapelle et al., 2008; Kane, 2006; Knoch & Chapelle, 2017). Test validity is defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" in the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 1999, p. 9). The interpretation of this definition has been widely accepted, but the establishment of the definition dates back to the early 1980s. Since long before then until the early 1950s, the centrality of interpretation of test validity had been criterion validity (Sawaki, 2011). Then in 1954, in *Technical Recommendations for Psychological Tests and Diagnostic Techniques*, the predecessor version of the professional standards for educational and psychological testing (AERA et al., 1954), test validity was defined with the following three types of validity, one of which has two subcategories: content validity, criterion-related validity with predictive and concurrent validities as subcategories, and construct validity (Messick, 1990, p. 7). This fragmented validity with various types of evidence evolved into a unitary perspective of validity that Messick (1989, 1990) proposed in the late 1980s. He claimed that test validity consists of a unitary concept and that construct validity is the basis of it (McNamara, 2006; Sawaki, 2011). What makes Messick's conception significant is his comprehensive interpretation of the validation process by integrating the consequences of test use into it (Kane, 2006; McNamara, 2006; Messick, 1990; Sawaki, 2011). The consequence in his concept, which is of interest to all stakeholders, led to the subsequent concept of test usefulness.

Bachman and Palmer's (1996, 2000) conception of test usefulness comprises the following six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. Moreover, it is worth noting that each quality should be weighted in a balanced manner to maximize the test usefulness. Therefore, Bachman and Palmer believed that which quality is prioritized should be flexibly adjusted according to the purpose and situation of the test. Despite the great advancement in the conceptualization of test validity, Messick's comprehensive concept as well as Bachman and Palmer's earlier concept lacked more practical guidelines for validation. Consequently, Bachman and Palmer's more accessible approach with specific guidelines were proposed. Still, Messick's key conception of the significance of consequences has been "influential on educational assessment and language testing in terms of broadening the scope of validation" (Im et al., 2019, p. 5). Thus, it has been passed down to the socio-cognitive model, the concept of "the consequential basis and fairness" (Weir, 2005, p. 5) in the field of measurement or assessment in education, and the AUA (Bachman & Palmer 2010).

Before discussing the structure of the AUA, a brief discussion of Toulmin model of argument (Toulmin, 2003) is in order. This model has been influential in many fields such as cognitive science, legal argumentation, and educational measurement (Kane, 2006; Mislevy, 1996). Toulmin's approach offers a framework to justify a statement or assertion with a systematic reasoning scheme with the following six components: claim (i.e., a statement, assertion, or conclusion), data (a fact that supports a statement), warrant (the legitimacy of data), backing (further evidence that supports warrant), rebuttal (limitation to a statement or unfavorable condition that challenges warrant/data), rebuttal backing (a fact that either supports or rejects/weakens a rebuttal), and counter claim (a conclusion that denies an original statement). With this framework, one can conclude whether his or her statement is justified or not justified step by step by following the practical reasoning scheme. Moreover, inclusion of more elements than a simple data–claim inferential link will make an argument stronger. Figure 1.1 illustrates the application of the quantitative part of the present study to Toulmin's model based on the AUA of Bachman and Palmer (2010, p. 97). The example illustrates how the claim is validated. A counter claim is in the dotted box because it is not supported in the example argument in the original model. When the rebuttal backing rejects a rebuttal, the original claim or hypothesis would be rejected, and a counter claim would be a conclusion instead.

Figure 1.1

The Toulmin Model Applied to the Present Study


Claim: Annotated diagram as writing feedback is helpful for students to understand coherence.
Counter Claim: Annotated diagram as writing feedback is **not** helpful for students to understand coherence.
Warrant: Students get better scores by revising coherence errors.
since
unles
Rebuttal: Students can revise their writing better without (annotated diagram) feedback.
so
on account
Reject
Backing: Students understand how to revise errors to make it coherent.
Data: Students' scores in the organization category improved in the second draft and the transfer task.
Rebuttal Backing: Scores with experimental group improved significantly better than control group.

Note. Based on Bachman & Palmer (2010, p. 97)

It should be mentioned here again that the Toulmin model has not only been embedded in the AUA on which the validation of this study relies; it has also been incorporated into classroom instruction as part of the argumentative paragraph writing approach. The Toulmin model approach has been the basis of the argumentation instruction of Common Core State Standards in the United States (O'Hallaron, 2014) and has been used in Japanese EFL writing classes to help improve persuasiveness in the construction of opinion statements (Sakamoto, 2016; Shimabayashi, 2019) as well as in L1 writing instruction in college (Suzuki et al., 2006). Toulmin's model of argument is a "well established means of understanding how to argue in one's first language, and it is hoped to help Japanese EFL students to understand the essence of argument as well" (Sakamoto, 2016, p. 12).

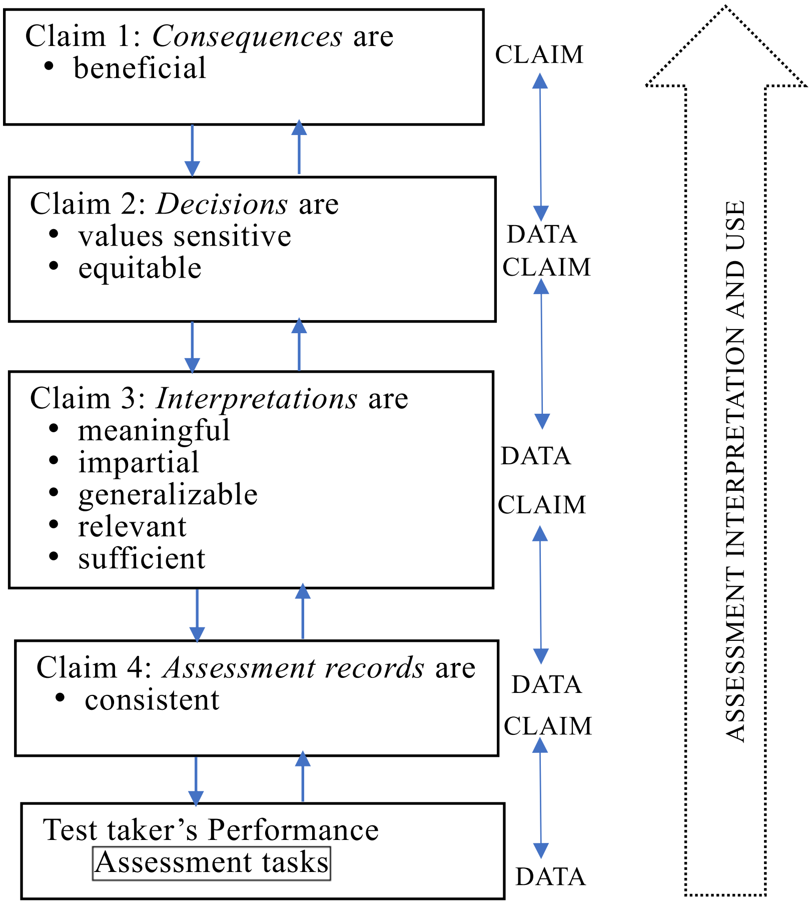
1.4.2 AUA Claims

Now, the AUA is discussed in terms of the structure of the framework with its components and the interrelationship across them. The AUA is defined as “a conceptual framework for guiding the development and use of a particular language assessment, including the interpretations and uses we (they) make on the basis of the assessment” (Bachman & Palmer, 2010, p. 99). As mentioned above, it follows Toulmin’s structure of practical reasoning.

The AUA includes four general claims (Claims 1-4) accompanied by warrants and rebuttals so that it can justify a particular assessment systematically based on the test taker’s performance. Claim 1 is related to consequence(s), Claim 2 is related to decision(s), Claim 3 is related to interpretation(s) about test taker’s language ability, and Claim 4 is related to the assessment record, such as a score and description. Each claim is supported by different types of warrants and challenged by rebuttals. The downward pointing arrow illustrates the development of language assessment, while the upward pointing arrow shows interpretation and use of assessment. All the related links are integrated by the author and illustrated in Figure 1.2 based on diagrams presented by Bachman and Palmer (2010, pp. 91, 100, and 104).

Figure 1.2

Inferential Links Across Claims Accompanied by Warrants and Rebuttals in Assessment Interpretation and Use



Note. Based on Bachman & Palmer (2010, pp. 91, 100, and 104).

The quantitative and qualitative assessments in this study are based on the AUA framework to evaluate validity assumptions regarding the use of an annotated diagram focusing on analysis of ideational and rhetorical coherence. The four claims based on the AUA examined in the present mixed methods research are presented below. The details on how the AUA is applied for the specific quantitative and qualitative sub-studies is presented in Chapter 3, as this information is closely related to the research methodology.

Claim 1: *Consequences* – The formative diagnostic assessment in an EFL writing classroom is beneficial for both the teacher and students in terms of teaching and learning (specifically in ideational and rhetorical coherence).

Claim 2: *Decisions*[3] – The diagnostic feedback comprises formative decisions that are made in classroom EFL writing assessment with the application of annotated diagrams are sensitive to or consistent with the existing values in terms of quality of coherence in writing.

Claim 3: *Interpretations* – The interpretations regarding the students' writing performance based on analytic criteria and annotated diagrams are meaningful.

Claim 4: *Assessment records* – The rating scale and text annotation used in the study function properly in the quantitative study, so the assessment records are obtained and analyzed appropriately and consistently.

1.4.3 Mixed Methods Design in This Study

With a mixed methods approach, an investigator gathers both quantitative and qualitative data, and draws interpretations by integrating the two to understand research problems (Creswell, 2015). Whereas quantitative data in language testing is clearly expressed in numerical form, such as scores on a rating scale, the form of expression of qualitative data is diverse. Qualitative research relies on nonnumerical data obtained from first-hand observation, interviews, questionnaires, focus groups, observations, etc. Qualitative methods include ethnography, grounded theory, discourse analysis, and interpretative phenomenological analysis (Creswell, 2012). In this study, in addition to

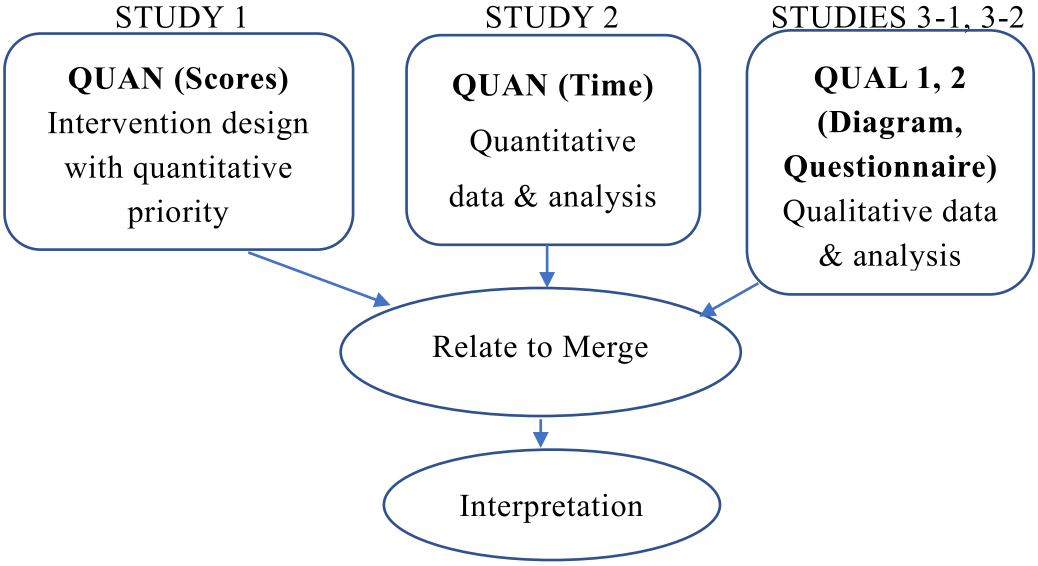
[3] Although the feedback is provided for students and teachers to make teaching/learning decisions as the classroom assessment, Claim 2 in terms of value sensitivity and equitability is not necessarily focused and is not included in the analysis in the present study.

the students' responses to the open-ended questions analyzed with thematic analysis, the visualized information obtained through an annotation tool is treated as qualitative data.

As stated in Section 1.1, the aim of the present study is to investigate the effectiveness of schematized diagnostic feedback that teachers generate by using an annotation tool on EFL argumentative stand-alone paragraph writing in the EFL writing classroom in a Japanese university. The present mixed methods research consists of three main studies: two quantitative studies and one qualitative study. To be more specific, it is a convergent mixed methods design using an intervention design. The convergent design of a mixed methods investigation is defined as the design involving the separate collection and analysis of quantitative and qualitative data, whose results are merged for data analysis (Creswell, 2015). With an intervention design, the researcher employs another design within a larger experimental framework. In the case of this study, integration consists of embedding the qualitative data of an annotated diagram as feedback within an experimental trial involving first draft, second draft, and transfer task (as a measure of retention) paragraph writing on argumentative topics. A simplified diagram of the overall flow is shown in Figure 1.3. A more detailed procedural diagram can be found in Chapter 3.

Figure 1. 3

Outline of the Mixed Methods Design of this Investigation


STUDY 1
QUAN (Scores)
Intervention design with quantitative priority
STUDY 2
QUAN (Time)
Quantitative data & analysis
STUDIES 3-1, 3-2
QUAL 1, 2 (Diagram, Questionnaire)
Qualitative data & analysis
Relate to Merge
Interpretation

Consistent with a multi-dimensional research design, this dissertation adopts a claim, which is a hypothesis, that needs to be validated at each stage of assessment. The findings are then succeeded by the next hypothesis, and it is validated. In a mixed methods approach, generating and testing hypotheses can be realized by various designs depending on the research purposes (e.g., Creswell, 2015; Kakai, 2015; Miller & Bustamente, 2016; Saville, 2016; Ziegler & Kang, 2016). Some researchers in Japan have focused on the development of a series of hypotheses and call it *Tadan-bunseki-ho* or the "multi-stage analysis method" or the "hypothesis-succeeding"[4] study (Yamanishi & Tanaka, 2003). These two approaches differ from each other in whether quantitative or qualitative analysis precedes hypothesis generation, while both attempt to combine

[4] The term and concept of "hypothesis-succeeding" was originally proposed by Saijo (2002), but his proposal was an attempt to succeed the hypothesis by repeating qualitative methods, while Yamanishi and Tanaka (2003) used both qualitative and quantitative data and analyses.

the advantages of the two analytical methods and complementarily capture the event of interest.

1.5 Overview of the Dissertation

This dissertation consists of six chapters. This chapter has offered an introduction to the research topic, the theoretical backgrounds for the present study, the validity framework used to articulate and evaluate claims and hypotheses to be validated, and the research design through which the investigation is carried out. Chapter 2 introduces the research assumptions relevant to the inquiry and reviews the literature on which this study is based. Chapter 3 discusses the research contexts, the methodological approaches, and the procedures. In Chapter 4, the results of the research questions (RQs) in both quantitative and qualitative studies are presented. Chapter 5 integrates and interprets the quantitative and qualitative results obtained in the previous chapters and discusses them. Finally, Chapter 6, the conclusion, summarizes key study findings, offer pedagogical implications and limitations, and suggests potential further research directions.

Chapter 2

Literature Review

Chapter 1 provided a research overview by introducing fundamental theoretical issues that underlie the research, including the argument-based validity framework. This chapter delves further into the theoretical underpinnings of the research by focusing specifically on the key issues on which the present study is based. The following five issues are covered in the chapter: the background of EFL writing instruction in Japan, formative assessment, coherence in writing, the effectiveness of writing feedback, text annotation tools, and graphic displays.

2.1 Current Issues of EFL Writing in Japan

2.1.1 Background of EFL Writing Instruction in Japan

First, it is necessary to discuss the issues surrounding EFL writing education in Japan to explore what is happening inside and outside of the classroom. The educational environment including school-based EFL education in the country is worth discussing as much as theory given the nature of this classroom-based study.

The results of national English language assessments by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) administered to 160,000 seniors in randomly selected upper secondary public schools in Japan for the 2013 and 2014 academic years were far from satisfactory to the stakeholders in English language education, although such results had been expected to some extent. The students' reading and listening abilities were in the high A1 and low A2 levels of the Council of European Framework of Reference (CEFR; Council of Europe, 2001). Even more disappointing was the finding that their speaking and writing performance levels were lower than their reading and listening levels: A majority of the students were classified at the A1 level. It is obvious that some measures had to be taken urgently to improve overall English language abilities in the country, and even more emphasis should be placed on productive skills such as speaking and writing. However, it must be noted that when it comes to the most high-stakes assessment of university entrance examination, there seems to be some argument especially among teachers in upper secondary schools regarding whether to incorporate the transitional trends into the university entrance examination system (Sawaki, 2017). Still, as the surveys in the following sections show, it is becoming a common understanding among the stakeholders in English language education that the development of logical writing skills is one of the most urgent areas to enhance the ability to communicate in the international community.

In response to calls to address the above-mentioned issue, MEXT has taken various measures in terms of development of English learners' productive ability in L2 including paragraph writing. The measures that MEXT has been taking are represented by an ongoing major government-initiated reform of English language instruction and assessment from the primary school to university levels (MEXT, 2018). In the following sections, EFL writing education in Japan is briefly discussed by exploring the situation in upper secondary schools and universities. Because this study is aimed at undergraduate students who are learning academic writing for the first time in college, it is necessary to explore EFL writing instruction before college.

2.1.2 Writing Instruction in Upper Secondary Schools

As mentioned above, there has been an increasing awareness of the significance of EFL argumentative performative skills, including the abilities to read articles critically and produce logically coherent passages. These skills are becoming more important than ever in secondary schools, which prepare students for higher education. Fundamentally, students in Japan are supposed to be taught paragraph writing with a basic knowledge of structure in an early stage of a formal academic writing course in university, but MEXT announced the new Course of Study (MEXT, 2018) to initiate learning that could form the basis of argumentative paragraph writing in high school beginning in 2022. The current subjects named *English Expression I, II* and *III*, which focus on English speaking and writing, are to be named *Logic & Expression I*, *II*, and *III*. As the name change indicates, there has been a major shift to focus more explicitly on improving the students' ability to produce language logically with the goals that "students express logically their opinions with reasons or grounds in detail in several well-structured paragraphs in English by processing expository passages or controversial issues" (MEXT, 2018, p. 173; translation by the author).

Two comprehensive surveys (Oi & Horne, 2016; Oi, Itatsu, & Horne, 2016) are notable because they have presented the situation of EFL writing instructions in secondary school shortly before the reform of the MEXT Course of Study. The surveys were conducted among EFL writing teachers and students in East Asia. The authors aimed to investigate writing practice, perception, and perspective before receiving writing instructions in college. The participants of the first survey (Oi & Horne, 2016) were undergraduate students who had been enrolled in the required EFL writing courses at national and private universities in Japan, South Korea, Taiwan, and Hong Kong (N = 1,447), with 779, 205, 254, and 119 students, respectively. Data were collected in 2012 in Japan and in 2014 in the other countries. The second survey's participants were junior and senior high school teachers in Japan, Korea, and Taiwan (N = 289) (Oi, Itatsu, & Horne, 2016).

The first study (Oi & Horne, 2016) showed some of the students' experiences with writing practices in the secondary school classroom before entering university. What the authors found was that students in high school EFL classes in Japan experienced exercises such as jumbled sentences or sentence-level translation from Japanese into English most frequently among the four countries/regions. Furthermore, the Japanese and South Korean participants had a "relative neglect or lack of frequency" of writing activities at length, such as in the form of a paragraph or an essay, in comparison with other countries/regions (p.92). The second survey (Oi, Itatsu, & Horne, 2016) targeted English teachers in secondary school concerning their instructions in classroom. They were asked what aspect of English writing they focused on in their writing instruction. For Japan, the greatest focus was placed on vocabulary and expression. Sentence writing and grammar were also important aspects for at least 44% (57/129) of the teachers, while only 15% (19/129) focused on paragraph writing and 7% (9/129) on essay writing. The sample size of these two independent student and teacher surveys might not be large enough to represent the country, but they consistently suggested that writing a coherent passage consisting of multiple sentences had not been among focal classroom activities, and that there were fewer opportunities to write paragraphs in secondary schools in Japan than in the other Asian countries studied. The results of insufficient writing instruction in secondary schools emerged in studies conducted by Kobayashi and Rinnert (2002, 2008), which also showed a lack of writing instruction up to high school in English.

Under the circumstances of EFL writing instruction in Japanese high schools described above, the revision of the Course of Study by MEXT as outlined at the beginning of this section was planned and implemented, and it had a significant impact on large-scale tests that measure English proficiency in Japan. The standardized tests were developed in Japan mainly for junior and senior high school students. Such tests including the EIKEN Test in Practical English Proficiency (the Eiken test, hereafter), the Test of English for Academic Purposes (TEAP), and the Global Test of English Communication for Students (GTEC*f*S) are sensitive enough to reflect MEXT's policy (2018) as they include independent writing or speaking tasks with argumentative topics in their test specifications. In particular, the Eiken test began to include an argumentative writing section in the Grade 2 test in 2016, and those in Grades Pre-2 and 3 in 2017 (Eiken Foundation of Japan, 2020). The Eiken tests are regarded as fairly high stakes tests in Japan, and they are often used for credit or qualification for admissions requirements in many high schools and universities. Thus, they carefully follow MEXT's guidelines and have been cautious about changes of their test specifications. Consequently, in the Eiken test, all the grades except the lowest two levels (Grades 4 and 5) now require opinion writing. This change has had a great impact on the stakeholders in secondary schools in Japan because taking large-scale standardized English tests often gives students a drive to improve their English proficiency, and the test format might be incorporated into English writing instructions in classrooms. Preparing for the Eiken test could offer an opportunity for students to write paragraphs (Tomita, 2019), which could be positive for the language assessments. The change in the test specification for the above-mentioned large-scale tests is, so to speak, a by-product, and MEXT's reform of the Course of Study has fundamentally intended to change instruction in the English classroom in secondary schools.

2.1.3 Writing Instruction in Universities

2.1.3.1 Target Language Use and Admission Tests in Japanese Universities. Before addressing the English writing instruction at universities specifically, it is necessary to discuss what English skills are required at the university level in Japan. In other words, the target language use (TLU) domain as EFL for writing classes at Japan universities, in this case, should be identified, because TLU tasks are within the defined TLU domain (Bachman & Palmer, 2010, p. 62). However, due to the myriad of majors and variable English proficiency at universities, the language skills required at Japanese universities are not always clearly defined. In other words, it is left to the discretion of individual universities, and there is no major national policy as a common understanding like the MEXT Course of Study up to high school. Furthermore, the format of the university entrance examination, which is the assessment that should essentially measure English language skills related to what EFL learners have learned so far and what they will be required to learn in the future in EFL classrooms at universities, does not seem to adequately measure the performance skills. It is true that a commercial standardized test, such as the TEAP, includes performance tests, and the results of the test are also utilized by some universities. However, the enforcement of a four-skills test for the Common University Entrance Examination has yet to be realized as of 2022,[5] despite the common understanding that the most urgent issue for English education in Japan is to strengthen performance skills and the fact that ongoing implementation has been considered for several years. While an admissions test plays a role as a measure of a student's achievement in secondary school EFL, "another important function would be to screen students who have sufficiently high English language ability to meet English language demands at the university" (Sawaki, 2017, p. 3). In addition to the issue of performance tests not being required content in entrance examinations, there is also the issue that the English language skills required by universities are not always clearly identified. As Sawaki (2017) remarks, needs analysis like the survey by Rosenfeld et al. (2001) would be desirable for the university entrance examination context in Japan to clarify the TLU domain definition, but those types of investigations seem limited. In the United States, Rosenfeld et al. (2001) identified frequently used tasks in both undergraduate and graduate courses in North America universities, where content instruction is conducted in English. Although the context is different from that in Japan, their findings are reliable and significant enough to be considered a prototype assessment task to provide support for the argument of the TLU domain and validity of the TOEFL iBT test (Chapelle et al., 2008; Sawaki, 2017). In fact, Green (2014) conducted a survey study in Japan with a large number of students and teachers in upper secondary school (*n* = 3868 and 423, respectively) as well as faculty members in Japanese private university (*n* = 19) to investigate the impact of introducing one of the four-skill assessments, the TEAP, to the country. They found that both high school teachers and students understand that the use of English in college requires a wider range of English skills, such as speaking and writing skills, than those

[5] Introduction of the four-skills assessment has been tentatively postponed to 2024 partly because the issue of fairness could not be resolved. Implementation after the year 2024 is being considered, but has not yet been finalized.

tested in the college entrance examination. Moreover, they generally believe that changes to admission procedures would bring about changes in the field of education and consider emphasis on the four skills to be desirable. Sawaki (2017) comments that his survey is valuable and supports the general direction of introducing four-skill assessment in Japan and could be supplemented with further studies. Changes in the content of university entrance examinations are expected to provide a positive feedback in terms of identifying the TLU domain in universities. Future trends in university entrance examinations should be monitored closely.

2.1.3.2 Writing Instructions in EFL Writing Classes and EFL Students' Perceptions of Their English Proficiency in Japanese Universities. Despite the difficulty summarizing English language instructions in Japanese universities, as noted in Section 2.1.3.1, there seems to be a common approach to writing classes in which college students learn EFL paragraph writing for the first time. Despite such limitations, the following two comprehensive surveys on instructions and classroom activities of EFL writing in Japanese universities provide some inferences about the writing instruction in the country. One is from another survey conducted by Oi, Horne, and Itatsu (2016) giving a questionnaire survey on instructors in university, both native English (NET) and non-native English teachers (NNET) (n = 17 and n = 23, respectively). The other survey conducted was conducted by MEXT (2021) used data from a general incorporated foundation, the Institutional Research Consortium of Japanese University, which consists of eight universities from both public and private in Japan with total of 55,624 participants. They reported college seniors' self-evaluation of their English proficiency level of four skills in 2020.

In the study conducted by Oi, Horne, and Itatsu (2016), regardless of whether they were NET or NNET, 83% of the teachers (33/40) required their students to work on argumentative writing tasks in their classes, and 63% of teachers (25/40) had their students write expository and descriptive tasks. They found that the most frequent feedback approach, regardless of whether the teachers were NET or NNET, was teacher feedback (40/40; 100%), followed by the process approach (35/40; 88%) and peer response (34/40; 85%), but less than half (19/40; 48%) responded that they used rubrics. Although the sample size is small, and thus there should be caution in generalizing the results, it is striking, but not surprising to the author, that nearly half of the instructors in the survey responded that they do not use a rating scale for evaluation. The author's personal impressions are consistent with this result, as the author knows that many EFL writing instructors often give a holistic rating of A, B, or C based on their own teaching experience. Experienced instructors may be able to provide consistent ratings and feedback based on evaluations without the use of rating scales, but it would be difficult to provide their students with a rationale for their ratings.

The study conducted by MEXT (2021) regarding the perception of English language proficiency showed that fourth-year college students had the lowest self-assessment of writing skills among the four skills. More specifically, 56.3% of the participants assumed that their English writing skills were at the CEFR A2 level or lower. Likewise, 55% said their English listening, 48% said their English speaking, and 37.3% said their English reading proficiency was at or below the CEFR A2 level. The undergraduates' lack of confidence in English writing is supported by another study conducted by the author (Matsumura, 2020), which includes factor analysis based on a questionnaire about English learning in an English class for academic purposes. The participants (*N* = 119) were first-year undergraduate English majors at a university in Tokyo. Among the four English skills, the students were the least confident in English writing and listening skills. Although the surveys conducted by MEXT and the author cannot simply be discussed in the same terms because of the difference in target population, sample size, and even context, it is expected that English writing skills are perceived as one of the most difficult English language skills throughout the university years. As Grabe and Kaplan (1996) state, writing skills do not come naturally and cannot be acquired without education and instruction in school or other settings.

Reports that Japanese university students lack confidence in their English writing skills emphasize the importance of discussing what kind of EFL writing skills should be developed at universities and what kind of instruction should be provided.

2.2 L2 Writing Feedback

The review of writing feedback is significant in the present study, whose objective is to investigate the effectiveness of teacher writing feedback. This section reviews the forms, focus, and content of the feedback returned to learners in response to writing assessment. This review provides insight into what is needed and what is missing to create formative diagnostic feedback for ideational and rhetorical coherence, the focus of this study.

2.2.1 Overview of the Trends of L2 Writing Instruction and Feedback

It is possible to review the trends of L2 writing instruction by looking at the transition of ESL writing instruction and the preceding L1 writing education in the United States. Oi (2004, pp. 69–70), consulting previous studies by Ferris and Hedgecock (1998), Grabe and Kaplan (1996), Horowitz (1986), and others, describes the development of L2 writing instruction in three stages: the form-oriented approach (since the 1960s), the writer-oriented approach (since the late 1970s), and the reader-oriented approach (since the late 1980s). She summarizes the three stages as follows. Until the 1960s, L1 English classes in the U.S. had focused on reading comprehension and writing as a means of analysis based primarily on reading literary works. In terms of writing activities, there was supposedly little intention of a systematic observation or instruction of the process of text creation, but the teacher commented on the final work created by the students following the rhetoric of the original model text. This was called the "product approach," which would contrast the so-called "process approach." The first phase of L2 writing instruction in the 1960s was similar to this L1 writing instruction. The form-oriented approach was considered to be primarily aimed at producing accurate sentences and establishing correct vocabulary and grammar. At the same time, mastery of rhetoric was another learning goal, and the "current-traditional rhetoric approach" was offered as a writing instruction in which model rhetoric was presented in advance. The second phase, which began in the late 1970s, is a writer-oriented approach, focusing on process writing, which is still at the core of writing instruction today. This approach involves multiple drafts, revisions, and edits before completing the product. Naturally, writing feedback plays an important role in the course of revision in the process writing approach. The third phase, which began in the late 1980s, is a reader-centered approach, in which the reader's point of view is emphasized, and the goal is to create writing with the reader in mind. ESL writing in the United States, especially in academic writing context, is assumed to require specialized rhetoric for the readers.

The summary above based on Oi (2004) illustrates the trends in L2 writing instruction at U.S. universities, and the global trends are similar. At the same time, the interpretation needs to consider the context at American universities in ESL environments. ESL learners at American universities are faced with the need to use English as a medium for input and output in a variety of subject areas, in the context of the so-called content-based approach. On the other hand, EFL learners at Japanese universities often learn English academic writing as a foreign language subject, except for learners involved in some English-medium instruction (EMI) courses. It should be kept in mind that the differences in the context can affect writing instruction.

Finally, when providing an overview on writing feedback, the Truscott–Ferris controversy cannot be ignored. It began when Truscott (1996) argued that grammar correction in L2 writing is ineffective. The abstract in his article begins with the rather provocative statement, "The paper argues that grammar correction in L2 writing classes should be abandoned for the following reasons" (Truscott, 1996, p. 327). The three reasons he cited are that its intriguing effects have not been demonstrated in research, that it is not expected to be effective in theory or in practice, and that it has rather harmful effects. This has been countered strenuously by Ferris and many other L2 writing practitioners (Chandler, 2004; Ferris, 1999, 2004; James, 1998). In fact, both sides of the argument make sense, but what is worth noting here is the specific reasons given by Truscott (1996) for the ineffectiveness of error correction. They are points that practitioners should definitely take into consideration when developing feedback: (a) There is a developmental sequence in the acquisition of grammar, and grammatical items pointed out in feedback can only be acquired at a stage appropriate to the learner; (b) the teacher's feedback corrections are not always consistent; and (c) excessive corrections may hinder the learner’s writing fluency and willingness to produce text. The second one, in particular, is a serious point regarding the reliability of the assessment. These perspectives are all important points to keep in mind when developing feedback.

Based on a meta-analysis of quantitative studies and related qualitative studies, Truscott (2007) concludes that error correction has “a small negative effect on learners’ ability to write” and that “if it has any actual benefits, they are very small” (p. 255). Ferris (2010), on the other hand, states that the discrepancy stems from the fact that writing feedback research is discussed from two different standpoints: L2 writing research and second language acquisition (SLA) research. That is, while L2 writing research focuses on the effects of feedback on learners’ text production improvement, SLA research focuses on the effects in terms of learners’ language development. Because it is not the purpose of this study, the differences between SLA and L2 writing perspectives will not be discussed in detail. However, the following process approach differences at the end are important as they relate to the research design of the study.

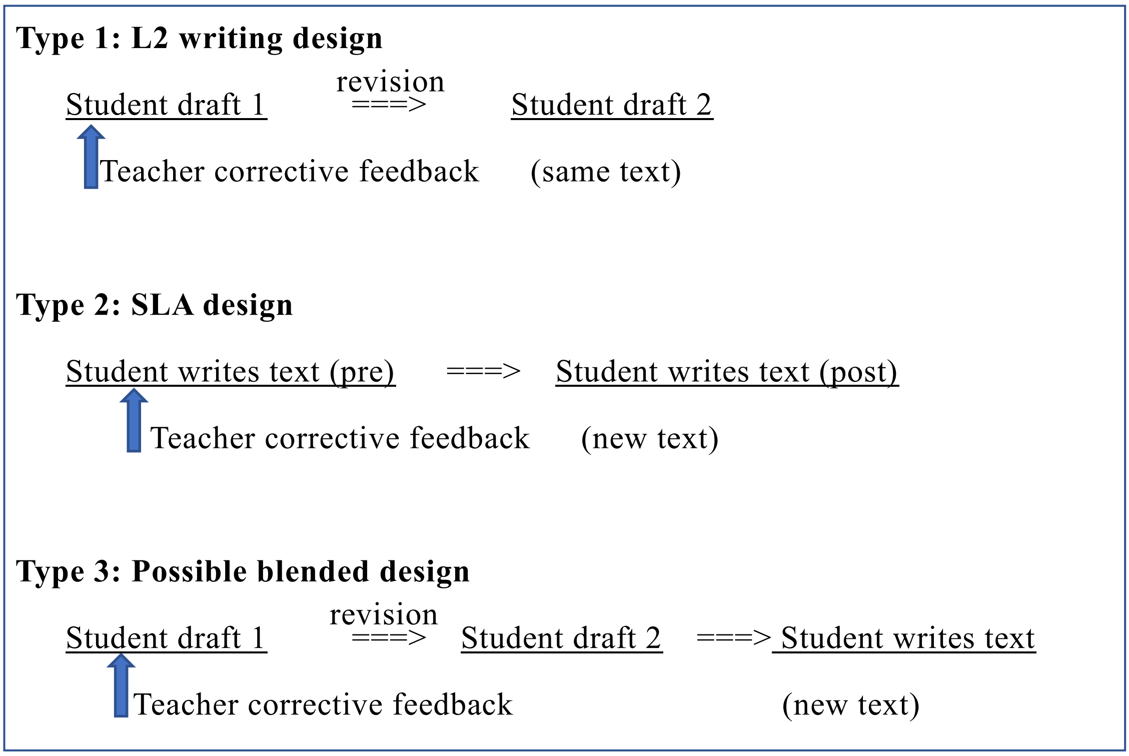
2.2.2 Types of Research Design Incorporating Revisions Based on Feedback

Ferris (2010) states that both L2 writing researchers and SLA researchers “often examine similar phenomena in similar ways,” but “they do not necessarily ask the same questions” (p. 181) in the study of written corrective feedback. While recognizing the differences in their viewpoints, she explores the points where they intersect and ways to take advantage of each other's strengths in actual research design. In terms of research design, the process writing approach of L2 writing classes, which includes response, revision, and subsequent textual analysis, is comparable to the experimental pretest, posttest, and delayed posttest designs of SLA research. She also illustrates three types of typical research designs incorporating revision of L2 writing, SLA, and a possible blended design.

Figure 2.1 shows three types of research designs, all of which include teacher corrective feedback and have the students (re)write repeatedly. These designs are based on the illustration presented by Ferris (2010, p. 195) with modifications by the author. Type 1 is an illustration of typical L2 writing research design, where students revise their drafts by referring to the teacher's feedback. Type 2 illustrates a typical SLA design, where students write text as the pretest and write new text as the posttest after being provided with teacher's feedback on texts produced during the pretest. Type 3 is presented as the blended research design, where students produce revised text in the L2 writing design in the first half of the session, and then write new text in the SLA design in the second half of the session. In seeking the intersection of the two different research approaches, she raises several issues to consider, one of which is "Do the effects of written corrective feedback (CF) in L2 writing classes endure beyond revisions of the same text to subsequent pieces of writing?" (Ferris, 2010, p. 197). The new type 3 blended design is proposed to address this inquiry.

Figure 2.1

Three Types of Research Designs Incorporating Revisions



Type 1: L2 writing design

Student draft 1 $\xrightarrow{\text{revision}}$ Student draft 2

Teacher corrective feedback (same text)

Type 2: SLA design

Student writes text (pre) ===> Student writes text (post)

Teacher corrective feedback (new text)

Type 3: Possible blended design

Student draft 1 $\xrightarrow{\text{revision}}$ Student draft 2 ===> Student writes text

Teacher corrective feedback (new text)

Note. The type numbers (1–3) have been assigned by the author for convenience. Adapted from Ferris (2010, p. 195).

Based on these previous studies, a modified design where teacher feedback is provided on the second draft prior to the transfer task from the language assessment point of view is employed in this study. Additional details are presented in Section 2.7.

2.2.3 Types of L2 Writing Feedback

Classroom writing feedback is defined as the return of evaluations and comments on work written by students. This may include self-evaluation, peer-evaluation, or teacher evaluation. As indicated in Chapter 1, this study deals with teacher evaluation and feedback. The way of giving feedback can also be diverse. The modality of teacher feedback can be verbal in the form of a conference, but written feedback is more common, in which descriptive comments are provided on the finished text or on a separate sheet. Furthermore, unlike feedback in speaking instruction, writing feedback is given based on the student's work after the text is created by the student; in this sense, all feedback is likely to be classified as delayed feedback. For a more specific example, Biber et al. (2011) conducted a comprehensive study with a meta-analysis on both L1 and L2 English writing feedback. Regarding the content of feedback, they show two types of feedback in case of L2 students only: feedback as comments and feedback with error identification in grammar/form rating accuracy as outcome focus.

Regarding what is given in teacher feedback, there are two main categories: those related to content and structure and those related to language use. While L1 writing feedback is likely to be primarily on content, especially in the case of an L1 report written by a college student with little or no language use problems (Tanaka, 2015), L2 writing feedback ordinarily includes both types of feedback[6]: grammar-based feedback, which identifies language-related errors in students written products, and content-based feedback related to content and structure-related issues. However, the weighting and presentation of both types vary depending on the purpose of the study and/or the learner's level of L2 proficiency.

Research from the late 1990s to the early 2000s explored the effectiveness of content-based feedback versus grammar-based feedback (e.g., Kepner, 1991; Oi et al., 2000; Sheppard, 1992) or which should be given first (Ashwell, 2000). For the former question, the authors concluded that content-based feedback is more effective than grammar-based feedback in improving the overall quality of writing. Among them, Oi et al. (2000), in their empirical study with a comparison group, reported that grammar-based feedback improved formality but decreased the total word count and lowered writing fluency. Furthermore, Ashwell (2000) compared three approaches in terms of

[6] These types of feedback have other labels. For example, Ashwell (2000) uses the terms form-focused feedback and content-focused feedback, while Oi et al. (2000) call them grammar-oriented feedback and content-oriented feedback, respectively.

the order of giving different type of feedback: the previously recommended approach of providing content-based feedback in the first draft and grammar-based feedback in the second draft (Zamel, 1985), the reverse order approach, and a mixed-type approach. He concluded that there were no significant differences among the three approaches.

Issues addressed by content-based feedback are presented in sentence form, mainly in the form of marginal comments or end-comments, whereas errors addressed by grammar-based feedback are pointed out with underlines, arrows, or other symbols. Table 2.1 shows a summary of types of writing feedback focusing on content-focused and form-focused feedback based on the previous studies (Biber et al., 2011; Bitcher & Storch, 2016; Storch, 2010; Tanaka, 2015).

Table 2.1

Content-Focused and Form-Focused Writing Feedback

Type of Feedback	**Target criteria**	**Form of presentation**	**Explicitness**
Content-focused feedback	Content, structure	Sentence (s) in marginal comments End comment	
Form-focused feedback	Grammar, mechanics, lexical choice	Underlines, circles, colors, codes, etc., often embedded in text	More explicit Direct feedback (error correction)
			Metalinguistic feedback*
			Less explicit Indirect feedback (underlines, circles, error codes etc.)

Note. Metalinguistic feedback provides evidence of errors, explanations of grammar, etc. (e.g., Lyster & Ranta, 1997; Tanaka, 2015).

Content-focused feedback is usually given in sentences in marginal or end comments (Ashwell, 2000, pp. 233, 235). Regarding grammar-based or form-focused feedback, there may be differences in the degree of explicitness of feedback, which can be broadly classified into two categories: direct and indirect (Ellis, 2010). Direct feedback refers to the provision of the correct form in the teacher response to the learner, while indirect feedback includes underlining errors or pointing out the type of error with an error code. In this case, the correct form is not presented, but the task of figuring it out is left to the learner. In both cases, it becomes more explicit when the rationale for the error is provided. The metalinguistic feedback in the "Explicitness" column in Table 2.1 can be given as additional information to the direct feedback as a grammatical explanation, or it can be provided to the student as indirect feedback that does not give the correct answer. Therefore, it has been placed in a position that covers both. In this sense, the former is more explicit if a meta-language explanation is given, and the latter is more explicit in given in the form of code rather than just underlining (Ferris, 2003; Tanaka, 2015). The written mode of feedback includes several forms such as marginal comments, end comments, editing codes, and circles/underlines (Biber et al., 2011).

In relation to the present study, in some respects it is difficult to provide direct corrective feedback to content and structure, but it is not clear how explicit it actually needs to be in order for the necessary modifications to be made. In fact, Ashwell (2000) gives a specific example as a comment provided as content-based feedback focusing on cohesion: "Think how to connect together the sentences in your first paragraph more smoothly" (p. 235). The actual effect of this feedback on learners deserves further study.

2.3 Diagnostic Language Assessment of EFL Writing

Because writing feedback in process writing involves evaluating students' written text for treatment, classroom writing activities that involve feedback are considered low-stakes assessment, requiring decisions about what to return to students and how and when to return it. Thus, they are considered to correspond to Claim 2, decision-making process, of the AUA (Bachman & Palmer, 2010). Moreover, corrective writing feedback diagnoses the students' writing for them to revise their drafts, so when appropriate and sufficiently detailed information is provided as a remedy, the evaluation process can be described as diagnostic.

The diagnosis of language proficiency is an old and new topic. The term diagnosis is often discussed in language education and applied to linguistics as well as diagnostic test in language assessment, but it is not necessarily well theorized or understood. Alderson's (2005) *Diagnosing Foreign Language Proficiency* is one of the most comprehensive books devoted to the topic of diagnostic assessment. He remarks that diagnosis in the language education field lacks exemplification or explanation, and the definition of the term is sometimes so superficial that they are just defined as illustration of strengths and weaknesses and their remediation. He claims this topic is under-researched and it is necessary to describe in detail what changes as learners develop, and more research should be conducted to support the descriptions. The insufficiency in the field has been ascribed partly to the dominance of high-stakes testing. Such tests are often conducted for norm-referenced assessment, where ranking-order of examinees is the main focus. However, he also claims that a diagnostic approach is also insufficient in the classroom context by saying, "even those who would concentrate their efforts on understanding classroom assessment procedures have failed to address the need for diagnosis of learners' strengths and weaknesses" (Alderson, 2005, p. 2). The diagnosis of language proficiency has long been discussed but may need to be explored anew by taking substantive approaches.

Table 2.2 presents a set of key phrases extracted by the author from the "hypothetical features of diagnostic tests" suggested by Alderson (2005, pp. 11–12). He explains that some of these features may contradict others as they are not necessarily definitive requirements for diagnostic tests but rather constituents of potential agenda for research. Thus, the list could offer guidance for further discussion and research on diagnostic assessment.

Table 2.2

A List of Key Phrases of the Features of a Diagnostic Test

	Key phrases of features of diagnostic tests
1	identify strengths and weaknesses
2	focus on weaknesses than on strengths
3	should lead to remediation
4	should enable a detailed analysis and report of responses to items or tasks
5	give detailed feedback
6	provide immediate results after test-taking
7	Typically, low-stakes or no-stakes
8	involve little anxiety to optimum performance
9	based on content which has been covered in instruction
or 10	based on some theory of language development, preferably a detailed theory
11	need to be informed by SLA research, or more broadly by applied linguistic theory
12	less "authentic" than proficiency tests
13	discrete-point than integrative, or more focused on specific elements
14	focus on language than on language skills
15	focus on "low-level" language skills than higher order skills
16	vocabulary knowledge and use are less likely to be useful
17	detailed grammatical knowledge and use are difficult to construct
18	language use skills like speaking, listening, reading, and writing are easier to construct than tests of language knowledge and use
19	enhanced by being computer-based.

Note. Adapted from descriptions given by Alderson (2005, pp. 12–13)

The features of diagnostic tests described in Table 2.2 could be summarized with several larger categories. The first is the category of is diagnostic test definitions and their significance. Diagnostic tests identify and report both strengths and weaknesses of learners' language knowledge as feedback, but identifying the weaknesses in particular encourages subsequent behavioral change of the learners and instructors. According to Alderson (2005), "the essence of a diagnostic test must be to provide meaningful information to users which they can understand and upon which they or their teachers can act" (p. 208). The second category is the granularity of information obtained by diagnostic tests. Diagnostic tests focus on more specific elements than on global abilities and more low-level skills than higher order skills, which enables providing detailed feedback for remediation. The third category includes the psychological aspects of the test and the effects they bring about. Because diagnostic tests are typically low- or no-stakes, which would lower affective barriers, they involve little anxiety that may lead to students' optimum performance. Last but not least, an important feature of diagnostic tests is immediacy. Providing immediate or slightly delayed results after test taking is valued. Giving learners immediate feedback on their performance is thought to have a maximum impact so that they can incorporate the feedback into their developing interlanguage. The importance of immediacy of feedback should be kept in test developers' or instructors' minds when conducting a diagnostic test. A computer-based test, especially one delivered over the Internet, is recommended by Alderson (2005) from the perspective of immediacy, which would allegedly enhance a diagnostic test. Hence, in his book he focuses on exploring the construct of DIALANG, the Internet-based language diagnostic test (https://dialangweb.lancaster.ac.uk/).

2.3.1 Direct and Indirect Writing Tests

DIALANG is an indirect type of language diagnostic testing system freely available on-line with five language skills or language knowledge (reading, listening, writing, vocabulary, and grammar, while speaking is excluded for logistical reasons) in 14 European languages. The test specification is based on the CEFR (Council of Europe, 2001), and the results are reported on the six levels of the CEFR (Alderson, 2005). DIALANG is an indirect test whose items assess knowledge without authentic application, and Alderson (2005) argues that it has been shown to be highly correlated with a direct test of writing. He even goes further and claims that using indirect test might be stronger than a direct test to identify relevant components of writing ability. However, as he acknowledges, the supposed existence of a high correlation between direct and indirect tests is becoming more doubtful among performance testing researchers. In reality, “indirect tests of writing are used less and less in this era of performance testing and therefore an argument can easily be made that diagnostic test of writing should be direct rather than indirect” (Knoch, 2007b, p. 4). Moreover, Knoch (2007b) suggests that “an indirect test for diagnostic tests lacks face validity and has fallen out of favor in general” (p. 14), which might lead to concerns among stakeholders as to whether the measurement results and content match. While there are obvious advantages to indirect testing, such as consistency in scoring and provision of immediate feedback, it goes without saying that it should be in line with the purpose of the assessment.

Direct/performance tests of writing, on the other hand, appear to ensure face validity that could be understood and used for instructions by stakeholders easily or convincingly. It should also be noted that there is an issue that they could provide sources of measurement error which arises from variability of tasks and rater judgements (Bachman et al., 1995). This may be two sides of the same coin with getting a wealth of information of useful sources of variation for diagnostic purposes. Even when considering the above-mentioned issues, writing performance tests have the advantage of being able to use rich and detailed text information for diagnostic purposes. However, it should be noted again that this is not such a major issue in the case of low-stakes or no-stakes assessments.

In this regard, Kunnan and Jang (2009) suggest the following in terms of test specification and test formats. As mentioned above, diagnosing language abilities requires detailed and careful analysis of test takers’ performance. To achieve that purpose, that is, to extract the elements of the test taker’s language ability while leaving out as little as possible, it is necessary to devise test specification and test formats carefully. Kunnan and Jang (2009) explain that among a range of test formats, performance-based assessments may be neither time efficient nor objective in scoring compared with other test formats such as multiple-choice response or fill in the blank, but they are considered to satisfy the needs to assess the students' achievement in the context of language use better in terms of second and foreign language processing in particular. A writing test follows one of the performance-based test formats that can provide a wealth of information for diagnostic feedback, where the construct of the rating scale should be carefully defined.

The rating scale and the descriptors are an important instrument of the diagnostic test, and they are closely related to what is being evaluated and what information can be extracted from the test results. Knoch (2011) discusses diagnostic assessment in terms of the development of rating scales. She argues, citing North (2003), that the rating scales are just a simplification of the test construct, but in effect they are often treated as test constructs themselves in the assessment. She claims that the theory and decision-making process on which the rating scale is based must be described to ensure its validity in development of the rating scale. Therefore, test developers may be required to show how a rating scale for a diagnostic assessment is constructed differently from that for a placement or a proficiency test. As Knoch (2011) argues, "rating scales developed for purposes other than diagnostic testing are not appropriate for diagnostic purposes" (p. 82). It may be possible to divert an existing rating scale to one for a diagnostic test, as Alderson (2005) also mentions. However, considering the need for more specific and detailed feedback in diagnostic assessment, she suggests that it is desirable that the rating scale be designed for that specific purpose.

The following sections discuss how coherence of EFL writing has been identified and evaluated, with a focus on the components and descriptors of the rating scales.

2.4 Coherence in Writing

Coherence in writing is the focus of this study; its nature has been described as tying together the entire text without detracting from the flow of the text based on the interpretation of other sentences. Watson Todd et al. (2007) remark on connectedness in discourse by noting that “connectedness refers to all the links, both explicit and implicit, in a text that make it a unified whole” (p. 11). They continue that connectedness is categorized into cohesion, which refers to explicit links, and coherence, which refers to implicit links. The implicitness in the latter would lead to the difficulty in being defined, taught, learned, and assessed. This implicitness has also prompted attempts to interpret its identity from multiple angles. In this section, various approaches in writing coherence as organization and connectedness in the discourse including progression of discourse are discussed.

2.4.1. L2 Organizational Pattern or Logical Development and Contrastive Rhetoric

Regarding organization of writing produced by EFL learners, the impact of cultural or ethnic backgrounds on rhetorical patterns has often been discussed. The contrastive rhetorical perspectives have traditionally contributed to explaining the EFL learners’ difficulty in writing argumentative paragraphs in English by claiming that it is attributable to logical differences in the approach between the two languages (e.g., Hinds, 1983, 1990; Kamimura, 1996; Kaplan, 1996; Miyake, 2007; Oi, 1984, 1999). Some authors have noted the negative effects of L1 on L2 by comparing the two languages and emphasizing the differences (Grabe & Kaplan, 1996; Kaplan, 1996). Kaplan’s (1996) description of the characteristic progression of paragraph organizations in Asian countries is a very well-known view: In the rhetoric of Asian nations, a passage keeps turning like a “widening gyre” (Kaplan, 1996, p. 10) and finally reaches a conclusion. There are many contrastive rhetorical studies for L1 Japanese writing. Hind’s (1983) examination of *tensei-jingo*, a daily newspaper essay column in the Asahi Shinmbun, is another well-known study among researchers in the field. Asahi Shinmbun’s reliable descriptions of quality are often used in university entrance examination questions of *kokugo* or Japanese as a national language. Hind (1983) points out the following as rhetorical characteristics observed in Japanese language: (1) a Japanese rhetorical pattern called *ki-sho-ten-ketsu*, which suggests that a discourse progresses by the following the four steps, that is, *ki* (introduction), *sho* (development), *ten* (twist), and ketsu (conclusion); (2) late introduction of the purpose of the passage called quasi-inductive; and (3) dependence of passage interpretations on readers, reader responsibility. Such a late appearance of the main claim or purpose of the passage or the progression of a passage from specific to general issue(s) is called inductive, which is often assumed to be a rhetorical pattern observed in the Japanese language. This pattern is also found in Japanese L2 writing (Kamimura, 1996; Oi, 1984), where, for example, topic sentences do not necessarily come early, but details come first in a narrative manner. On the other hand, a deductive rhetorical pattern, where a passage progresses from general to specific issues, is observed in English L1 writing because in English (or Western) rhetoric (Hinds, 1983, 1990), the general framework or conclusion is presented first, followed by specific examples and explanations.

As a related study, Oi (1999) investigated EFL argumentative writing produced by Japanese L1 undergraduates (*n* = 32) by comparing against English L1 counterparts' writing (*n* = 33). The main findings include Japanese L1 students' hesitance in making a claim, indecisiveness in reaching a claim, occasional inconsistency about the claim, as well as lack of support. However, she suggests that, although these characteristics could be attributed to contrastive rhetorical issues due to cultural difference, it is also a matter of education and that it could be addressed by learners' learning academic rhetoric specific to English.

Kubota's (1998) study with Japanese college students with little or no experience of staying in English-speaking countries (*n* = 46) revealed that half of the students used different rhetorical structure for L1 writing (Japanese) and L2 writing (English). She suggests that, as their writing either in L1 or L2 is different, it may not be possible to treat them as one group, much less to ascribe their difficulty in L2 writing to one factor, that is, a contrastive cultural issue. In other words, this study could be interpreted from different perspectives. As Yamashita (2019) suggests, L1 organizational patterns do not necessarily negatively transfer to L2. Hirose (2003) conducted a similar study with a relatively small sample size (*n* = 15) on organization patterns of Japanese English learners' argumentative writing. The results showed deductive organizational pattern were used in all L2 writing and in some L1 writing. The L1 and L2 organization scores were not significantly correlated, and the L2 total score was not correlated with the L1 total score. Moreover, the L2 organization and total scores were significantly different from those of L1 writing. Based on these results, she suggests that EFL argumentative writing by Japanese college students does not always follow the inductive organization pattern, and it is not necessarily appropriate to centrally associate the L2 and L1 organization. Hirose (2005) confirmed the above results and concluded that the factors explaining organizational difficulties of EFL students whose first language is Japanese are diverse and that they should not be explained in terms of cultural difference in discourse solely. The results shown in the studies cited above have led to an argument that some other variables such as the writer's L2 proficiency, writing experience in L2, and the writing instructions they have received should be taken into consideration besides Kaplan's contrastive rhetorical factors, which should have some influence on their L2 organization (e.g., Connor, 1996; Matsuda, 1997).

Lastly, Sasaki and Hirose (1996) explored factors affecting the L2 writing rhetorical pattern in the Japanese L1 context, which is interpreted as a synonym for organization or coherence in a broad sense. They investigated factors that might influence Japanese university students' expository writing in English (*n* = 70) along a variety of dimensions: L2 proficiency, L1 writing ability, writing strategies in L1 and L2, metaknowledge of L2 expository writing, past writing experiences, and instructional background in quantitative and qualitative approaches. Their quantitative analysis revealed that (a) students' L2 proficiency, L1 writing ability, and metaknowledge were all significant in explaining the L2 writing ability variance; (b) among these three independent variables, L2 proficiency explained the largest portion (52%) of L2 writing ability variance, followed by L1 writing ability (18%) and metaknowledge (11%); and

(c) there were significant correlations among these independent variables (Sasaki & Hirose, 1996, pp. 137–138). The qualitative analysis revealed that good writers were significantly different from weak writers in the degree of attention to overall organization while writing in L1 and L2, fluency in L1 and L2, the degree of confidence in L2 writing for academic purposes, and the frequency of writing English paragraphs while in high school. However, there was no significant difference between good and weak writers for the other writing strategies.

As the aforementioned studies have indicated, although it is certain that the EFL writing produced by Japanese L1 students is influenced by the rhetorical pattern of their first language, it is inappropriate to explain the difficulty of constructing organized writing in English by contrastive rhetorical perspectives solely because other factors seem to have a greater influence. Moreover, as Oi (1999) suggests, the difficulty should be considered a matter of education; thus, it is necessary to identify the nature of English organization from various perspectives to offer meaningful instructions to help students construct strong and coherent English writing.

This section has discussed the L2 writing structure of Japanese EFL learners mainly in terms of contrastive rhetoric. The issue of organization in L2 writing for Japanese university students, however, is not only a question of rhetoric or logical development of inductive or deductive; it is also a matter of coherence or cohesiveness of the text. The studies have consistently suggested the lack or insufficiency of coherence in Japanese college students' writing, which leads to disorganization of discourse (Harder & Kutz-Harder, 1982; Hinds, 1987; Nishigaki et al., 2007; Yamashita, 2019). Moreover, there are several previous studies on coherence of L2 English writing by Japanese college students in particular, and they suggest that teachers' writing feedback does not cover organization or coherence, and as a result, students' writing often remain less logical (Tsuji, 2016; Yasuda, 2006; Yasuda et al. 2014). Therefore, in the next section, the rating scales focusing on organization and coherence are discussed to explore the conceptual definitions of coherence underlying the rating scales.

2.4.2 Conceptual Definitions of Coherence Through Descriptors in the Language Assessment Framework

In this section, the descriptors or wording representing organizational aspects of writing are explored to examine how they are defined in the existing rating scales. This overview is necessary given the vague nature of coherence (Cerniglia et al., 1990; Knoch, 2007b), which may result in difficulty providing clear descriptors or organization/coherence in rating scales and effective writing organizational assessment. Comprehensive investigation of the existing rating scales gives some information about the overview of the position of organizational aspects in rating scales. Wagner (2015) reviewed a wide range of the writing rating scales with the sub-skills and compiled a summary.[7] As she "attempted to be true to the terms used in the original scales," (p. 32) the terms such as *cohesion* and/or *coherence* as skill(s) are used in the place of organization in some scales, while they are combined into one subcategory in other scales. Her summary includes both holistic and analytic scales. For holistic scales, she examined descriptors for specification of constituents of writing skills. The 39 sources were primarily based on rating scales used in higher education contexts including major standardized tests. She tallied the frequency of each subskill used in the scales, whose result showed the predominance of five subskills: content/ideas, grammar, organization, vocabulary, and mechanics (*n* = 31, 26, 26, 26, and 20, respectively), which agree with the conventional writing subskills in general. Organization is a criterion label used in many rating scales as a collective term referring to rhetorical elements including coherence and cohesion. What should be noted here is that organizational aspects in

[7] The writing scales subskills of the survey by Wagner (2015) comprised 21 studies, including the predominant five subskills cited in the text. The organization-related subcategories are rhetorical features (*n* = 3) and coherence/cohesion (*n* = 11). Refer to Wagner (2015, pp. 251–255) for more details.

rating are not a privilege of analytic scoring. They are also included in holistic scales as descriptors, even though they are not constructed as independent subcategory. In writing evaluation, the organization and coherence of writing is an indispensable component of any rating scale.

As an illustrative example, Table 2.3 shows the descriptors of thematic development and coherence presented in pragmatic competence in the CEFR descriptors (2018). Pragmatic competence is defined as follows: the user/learner has knowledge to produce messages with (a) discourse competence, (b) functional competence, and (c) design competence. Among them, discourse competence is most related to writing coherence, where the user/learner should be able to produce messages that are organized, structured and arranged. The author extracted the relevant descriptors to writing competence presented in Table 2.3 because the original table in the CEFR descriptors also includes statements for speaking competences. The author has paraphrased and simplified the descriptions for convenience. The two categories of thematic development, and coherence and cohesion out of the six pragmatic competence presented in the CEFR descriptors are shown in the table. Pragmatic competence illustrated in the CEFR descriptors has six categories: flexibility, taking the floor, thematic development, coherence, propositional precision, and spoken fluency.

Moreover, in the context of CEFR, the *Self-assessment Grid for Overall Written Production*, is presented in Table 2.4. This would provide an overall assessment of the learners' written products across the CEFR levels.

Table 2.3

Descriptors of Discourse Competences of Written Products in CEFR Pragmatic Perspectives (Based on CEFR Companion Volume. 2018, pp. 141-142)

Competence	PreA	A1	A2		B1		B2		C1	C2
CEFR: Thematic Development	*No descriptor available*		*No descriptor available*	・can give an example in a simple text with 'like' or 'for example'	・shows awareness of the conventional structure of the text type concerned	can clealy signal chronological sequence in narrative text	・can develop a clear argument with supports and relevant examples ・can evaluate advantages an disadvantages	・can present and respond to complex lines of argument convincingly	・can write a suitable introdcution and conclusion to a longer text ・can hold a target reader's attention	・can use the conventions to communicate effectively ・can fulfill all the communicative purposes
CEFR: Coherence and cohesion	*No descriptor available*	・can link with linear connectors like 'and' or 'then'.	・can link with simple connectors like 'and', 'but' and 'because'	・can link simple sentences to tell a story ・can describe something as a simple list of points	・can link sentences with cohesive devices ・can make simple logical paragraph breaks	・can intorduce a counter argument (with however)	・can produce text generally well-organized and coherent ・can structure longer texts in clear, logical paragraphs	・can use a variety of liking words efficiently to mark clearly the relationships between ideas	・can produce well organized, coherent text ・a variety of cohesive decices and organizational patterns	・can create coherent and cohesive text ・can make full and appropriate use of a variety of organizational patterns ・can use a wide range of cohesive devices

Note. Relevant items to written products were extracted exclusively and the descriptors were paraphrased in more simplified form by the author.

Table 2.4

Descriptors of Self-assessment Grid in CEFR (Based on CEFR Companion Volume. 2018, p. 169)

Production	A1	A2	B1	B2	C1	C2
CEFR: Overall Written Production	･can write simple isolated phrases and sentences	･can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but' and 'because'	･can write straightforward connected texts on familiar topics or of personal interest	･clear, detailed texts on a variety of his/her field of interest ･can write an essay or report with reasons or agains a particular view point	･can express oneself in clear, well- structured texts at some length ･can write detailed expositions of complex subject ･can write different kinds of texts in a style appropriate to the reader in mind	･can write clear, smoothly-flowing text ･can write with an effective logical structure which helps the reader to find significant points ･can write summairies or reviews of literary work

Note. The descriptors were paraphrased in a simplified form by the author.

In A1, connection between sentences is not expected because students are required to complete isolated sentences. In A2, the user/learner is supposed to be aware of the connections between adjacent sentences such as and, but, like, and for example. In B1, the user is required to arrange text units in proper order in chronological writing, which requires coherence skills. In addition, logical development is required to relate the point of view of others in opinion statements. In B2, the emphasis is on content and quality, and the writer must have a higher degree of coherence to be persuasive and effective and must be able to organize his or her writing in a larger framework that spans multiple paragraphs. In C1 and C2, students are required to weave coherent text in academic writing with a specific audience in their field of specialization in mind.

Next, to examine what descriptors are included in the rating scales of standardized tests in Japan, three representative large-scale English language proficiency tests in Japan are reviewed. Table 2.5 shows descriptors of independent writing tasks in the three large-scale English proficiency tests in Japan: the Eiken test (Eiken, 2016), the GTEC*f*S (Benesse, 2021), and the TEAP (Eiken, 2019). A correspondence chart comparing each test to the CEFR is provided as a guide, though not an absolute one. The Eiken test does not officially disclose the assessment grid for organization, except for the differences in prompts for each level and the total number of words required, among other factors. The descriptors are basically the same for each level, although they are expressed slightly differently as noted in the two broad categories of the upper three levels—Grades 1, Pre-1, and 2—and the lower two levels—Grades Pre-2 and 3.

Table 2.5

Descriptors of Organizational Skills in Specific in Independent Writing Tasks in the Three Large-scale English Proficiency Tests in Japan

CEFR level	A1	A2	B1	B2	C1	C2
Test	Grade 3	Grade pre-2	Grade 2	Grade pre-1	Grade 1	—
Eiken Test（Eiken, 2016）	・The structure and flow of English sentences are easy to understand and logical ・You can effectively use expressions (conjunctions, etc.) that indicate the flow and development of the information you want to convey, making it easier to understand your own opinions, their reasons, and the overall structure of the English text.					—
	0 0.5-1.5	2-3.5	4-5.5	8-6		—
GTEC *f*S (Benesse, 2021)	Words and sentences are written, but there is no connection between the contents.	Some ideas are written, but it is difficult to see the connection of the contents	There are parts where it is difficult to see the connection of ideas, but the content is somewhat cohesive.	The connection of the written contents is easy to understand, and it is well organized as a whole.	—	—
	33-26	59-34	87-60	98-88	100-99	—
TEAP (Eiken, 2019)	Unrelated to task/topic. Fewer than 50 words. Copied directly from the input text with little or no original language. (Below A1:25-20)	(Coherence) No logical paragraph structure or some separation which is not appropriate; text consists of mainly unconnected sentences with no clear direction or progression across sentences. (Cohesion) Uses conjunctions to link clauses within sentences, but generally does not mark clearly the relationship between sentences. Use of referential cohesive devices (for example, pronominal reference) is generally not clear.	(Coherence) Has a logical structure but the organization of ideas may not always be clear; organized into paragraphs, but the paragraph structure may not be completely appropriate. (Cohesion) Sentences and paragraphs are generally connected using discourse markers; use of referential cohesive devices (for example, pronominal reference) is mostly clear.	(Coherence) Organized as a coherent response to the task; organization of ideas within and across paragraphs is generally clear, though may be formulaic. (Chesion) Uses discourse markers and referential cohesive devices effectively to mark the relationship between sentences and link utterances into clear, coherent discourse.		—

As can be seen in Table 2.5, there is no descriptor of C2 for any of the tests and there is no description of C1 in GTEC*f*S, which suggests that students at those levels are not the target. Interestingly, in the scales in Table 2.5, the lower-level descriptors are more detailed, while in the CEFR descriptors in Table 2.3, the higher levels are more detailed. This is probably due to the fact that the target population for the test is Japanese EFL learners, the majority of whom are distributed in the middle to lower CEFR levels, and that more detailed descriptors are needed at those level. It indicates the importance of establishing descriptors at a granularity that matches the target population.

There are two points to be note in Table 2.5. First, the Eiken test's descriptors for the writing test are not simple and do not show clear distinction across the levels. This might be due to the relatively new introduction of the writing test or Eiken's policy, but it would not be helpful to the test taker's learning. On the other hand, the TEAP includes detailed descriptors of organization as well as other subskills. This may be due to the fact that TEAP has been developed assuming the introduction of four skills for university entrance exams, so there is a greater need for accountability for assessment.

2.4.3 Conceptual Definitions of Coherence in L2 Writing Studies

In the above sections, writing coherence has been reviewed through descriptors of how coherence is addressed in the CEFR and in standardized language tests in the Japanese context. To review how L2 English writing researchers have defined writing coherence, this section focuses on the following studies: (a) metadiscourse connectors (Crismore et al., 1993); (b) coherence in topical structure analysis (TSA) (Knoch, 2007b); (c) development of checklists for diagnostic assessment (Kim, 2011); (d) taxonomy of knowledge-based academic writing skills (Grabe & Kaplan, 1996); and (e) defining L2 writing components through raters' decision-making behaviors (Cumming et al., 2001, 2002).

2.4.3.1 Metadiscourse Connectors. Crismore et al. (1993) and Intaraprawat and Steffensen (1995) propose analyzing coherence by using metadiscourse markers. According to Crismore et al. (1993), metadiscourse refers to the writer's discourse about their discourse, where his/her directions for how readers should read, react to, and evaluate what he/she has written about the subject matter (p. 39). They argue that the writer guides and directs readers behind the text to help them organize and understand the content embedded in the form of text, so that they can better appreciate the writer's intention or attitude toward the text.

In their study on metadiscourse of texts written by American and Finnish university students on argumentative topics, Crismore et al. (1993) used writing with a persuasive aim because writers are more likely to use metadiscourse in this type of writing (Williams, 1989). Moreover, the possibility of using metadiscourse would contribute to increase the quality of persuasive writing by students (Cheng & Steffensen, 1996), and thus EFL learners would benefit from explicit teaching of metadiscourse of the targeted language (Mauranen, 1993). These benefits of using persuasive writing can also be theoretically supported by the idea of schematic structures of discourse, which refer to genres of the text proposed by Kintsch and van Dijk (1978). Argumentative/persuasive writing is a typical example of conventional schematic structures of discourse, which helps readers to "understand discourse as a story" (Kintsch & van Dijk, 1978, p. 366). The use of metadiscourse would be helpful in this process.

Crismore et al. (1993) adopted Vande Koppels' (1985) system of classification of metadiscourse with modification. They kept Vande Koppels' two major categories intact but revised the system by collapsing and reorganizing their subcategories. Table 2.6 summarizes the metadiscourse categories with examples proposed by Crismore et al. (1993). They are broadly grouped into two major categories: textual metadiscourse and interpersonal metadiscourse, each of which has several subcategories. Specific examples are provided for each category. Their presentation of the concept and specific use of metadiscourse reminds us once again that the use of connectors is a technique used selectively by a writer as a guiding beacon for the reader.

Table 2.6

Classification System for Metadiscourse Categories

I. Textual Metadiscourse	
1. Textual markers	Examples
(a) Logical connectives	*and*, *but*, *therefore*, *in addition*, etc.
(b) Sequencers	Numbering words like *first*, *second*, etc.
(c) Reminders	*As I mentioned earlier*, etc.
(d) Topicalizers	Indicating topic shift like *now*, *in regard to*, *speaking of*, etc.
2. Interpretive markers	
(a) Cod glosses	*for example*, *what I mean is*, etc.
(b) Illocution markers	*I state again that…*, *to sum up*, *to conclude*, etc.
II. Interpersonal metadiscourse	
3. Hedges	Modal auxiliaries like *can*, *could*, *may*, *and might* in epistemic readings; verbs of cognition with a first-person subject like *I think*, *I feel*, *I guess*, *I suppose*; uncertain adverbs like *perhaps*, *maybe*
4. Certainty markers	*I am absolutely sure that…*, *It is clear that…*
5. Attributors	Indicating the source of textual information like *Einstein claimed that…*
6. Attitude markers	*Hopefully*, *doubtfully*, *unfortunately*, *most importantly*, etc.
7. Commentary	*You may not agree that…*, *think about it* Following expressions like *you* as directing the readers; real questions which would be answered by the writers later on; tag questions, etc.

Note. Adapted from Crismore et al. (1993, p. 47).

2.4.3.2 Coherence in TSA. The quality of text coherence is evaluated through a careful reading of a written product by examining connections between sentences and the development of the topic in a paragraph or as a whole passage. It is necessary to explore whether the reader can follow the text without losing the thread of the argument while reading. In this section, several studies on local coherence, which refers to the connection between adjacent sentences based on propositional succession, are reviewed.

As described in Chapter 1, text coherence is propositional coherence based on the propositional content of discourse (Lautamatti, 1990). As such, it is defined as the "semantic property of discourses, based on the interpretation of other sentences" (van Dijk, 1977, p. 93). In linguistic interpretation, a proposition is the core meaning of a clause or a sentence that is constant: It is not changed by the grammatical voice or exceptional sentence patterns.[8] Therefore, a proposition is the essence or the content in the sentence, which is expressed with theme (topic/subject) and rheme (comment/predicate). This concept can be explained with topical structure analysis (TSA), which originated from the functional sentence perspective (FSP) (Daneš, 1974).

This section focuses on the outline of TSA by featuring the study in which Knoch (2007b) developed the rating scale for coherence. TSA was first described with the intention of analyzing topic development in reading material in the context of text readability based on topic and comment, or theme and rheme, based on the analysis by Lautamatti (Knoch, 2007b). Lautamatti (1987) defined the

[8] As criteria for identifying sentence topics, the following sentence patterns are exceptions: cleft sentences, the anticipatory pronoun "it", the existential "there", and introductory phrases (Wu, 1997, p. 57). See Wu (1997) for more details.

theme of a sentence as “what the sentence is about” and the rheme is the comment of a sentence or “what is said about the theme.” Daneš (1974) originally proposed the thematic progression (TP) theory based on the FSP on the framework of theme and rheme. The TP theory has much in common with TSA: They share the same basic concept because both Lautamatti and Daneš belong to the Prague School of Linguistics. Lautamatti’s three types of progression of TSA summarized by Hoenishch (1996) are introduced by Knoch (2007b), namely parallel progression, sequential progression, and extended parallel progression. In addition to these three types, Schneider and Connor (1990) suggest dividing sequential progression into three subcategories: direct sequential progression, indirect sequential progression, and unrelated sequential progression.

Based on the six progression categories suggested by Schneider and Connor (1990), Knoch (2007a) investigated 602 written scripts in a pilot study using the rating scale of DELNA, which is administered at the University of Auckland, New Zealand. The pilot study led to creating two more categories. One is specifically for ESL candidates whose English proficiency levels are relatively low: coherence break, which means “attempt at coherence fails because of an error.” The other is superstructure, where “coherence is created by a linking device instead of topic progression” (Knoch, 2007a, p.115). Table 2.7 provides a comparison of progression types among TP, Lautamatti’s, Schneider and Connor’s, and Knoch’s TSA.

Table 2.7

A Comparison of Progression Types Among TP, and Lautamatti's, Schneider and Connor's, and Knoch's TSA

TP_Daneš (1974)	TSA_Lautamatti (1987)	Schneider & Connor's (1990)	Knoch (2007a, 2007b)
Constant T.P. <a, b> <a,c> <a, d>	*Parallel progression* <a, b> <a, c> <a, d>	*Parallel progression* <a, b> <a, c> <a, d>	*Parallel progression* <a, b> <a, c>
Simple Linear T.P. <a, b> <b, c> <c, d>	*Sequential Progression* <a, b> <b, c> <c, d>	(*Sequential Progression*) *Directly Related Sequential progression* <a, b> <b, c> <c, d>	*Directly Sequential progression* <a, b> <b, c>
Derived T.P. <a, b> <hypertheme,c,d> <hypertheme, e,f>	*Extended Progression* <a, b> <b, c> <a, d>	*(Sequential Progression)* *Indirect Related Sequential Progression* <a, b> <indirect a, c> or <a, b> <indirect b, c>	*Indirect Related Sequential Progression* <a, b> <indirect a, c> or <a, b> <indirect b, c>
			Superstructure Coherence is created by linking device instead of topic progression <a,b><liking device, c,d>
		(Sequential Progression) *Unrelated Sequential Progression* <a, b> <c, d>	*Unrelated progression* <a,b><c,d>
		Extended Progression <a, b> … <a, c> or <a, b> … <b, c>	*Extended Progression* <a, b> … <a, c> or <a, b> … <b, c>
			Coherence break Attempt at coherence fails because of an error <a,b><failed attmpts at a or b or linker,c>

Note. Adapted from Daneš (1974), Knoch (2007a, 2007b), Lautamatti (1987), and Schneider and Connor (1990).

Table 2.8 shows examples of each progression presented by Knoch (2007b, p. 124). Some of the examples are from the author of the present study.

Table 2.8

Examples of TSA Categories

TSA categories and examples	**Definition by symbols**
1.Paralel progression My mother's kitchen is old and small. It is, however, nice and cozy*.	<a, b><a, c>
2. Direct sequential progression The report shows some differences in eating habit of men and women. These differences include an age factor*.	<a, b><b, c>
3. Indirect progression The main reasons for the increase in the number of immigrates is the development of some third world countries, e.g., China. People in those countries have enough money to support themselves living in a foreign country.	<a, b><indirect a, c> or <a, b><indirect b, c>
4. Super structure The popularity of Japanese animation has moved to the next stage. For example, *Demon Slayer*, which is R-rated was accepted as a movie for adults*.	<a, b><linking device, c, d>
5. Extended progression The first line graph shows New Zealanders arriving in and departing from New Zealand between 2000 and 2002. The horizontal axis shows the times and the vertical axis shows the number of passengers which are New Zealanders. The number of New Zealanders leaving and arriving has increased slowly from 2000 to 2002.	<a, b>…<a, c> or <a, b>…<a, c>
6. Coherence break All the animals in the zoo looked unhappy to me. It is in a cage all day*.	<a, b><failed attempts at a or b or liner, c>
7. Unrelated progression I live in the dormitory in college. Tokyo is the biggest city in Japan*.	<a, b><c, d>

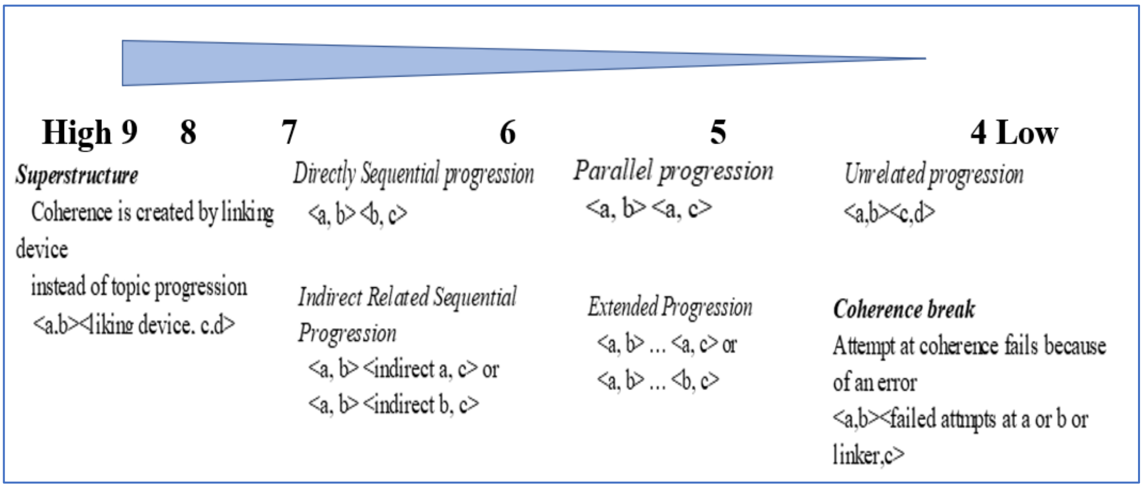
Note. Adapted from Knoch (2007b, p. 124). An asterisked sentence (*) indicates an example from the author.

In Knoch's main analysis, she attempted to identify which categories were used by students at different proficiency levels by correlating the DELNA final scores with the percentage of occurrence of each category. Based on the correlations between TSA variables and the DELNA final scores and the box plots for each of the variables, she examined the quantitative results to develop a TSA-based rating scale for coherence with five levels (4 through 9). The descriptors for each level are as follows. Level 4 is defined as "frequent unrelated progression and coherence breaks" are observed, whereas "sequential progression, superstructure, and indirect progression" are identified infrequently. Level 5 is defined as "as level 4, but coherence might be achieved in stretches of discourse by overusing parallel progression," and "only some coherence breaks" are identified. Level 6 is defined as a "mixture of most categories" are observed, "superstructure" is "relatively rare, " and there are "few coherence breaks". Level 7 is defined as "sequential progression" is observed frequently, "superstructure" occurs more frequently, but "parallel progression" is observed infrequently, and "possibly no coherence breaks" are identified. Level 8–9 is defined as a "writer makes regular use of superstructures, sequential progression", and "few incidences of unrelated progression" or "no coherence breaks" are identified (Knoch, 2007b, p. 180).

Figure 2.2 shows the relationship between rating level and progression type created by the author based on the above description and the rating scale of Knoch (2007b, p. 180).

Figure 2.2

Relationship Between the Rating Level and TSA Progression Types


High 9
8
7
6
5
4 Low
Superstructure
Coherence is created by linking device
instead of topic progression
<a.b><liking device. c.d>
Directly Sequential progression
<a, b> <b, c>
Indirect Related Sequential Progression
<a, b> <indirect a, c> or
<a, b> <indirect b, c>
Parallel progression
<a, b> <a, c>
Extended Progression
<a, b> … <a, c> or
<a, b> … <b, c>
Unrelated progression
<a,b><c,d>
Coherence break
Attempt at coherence fails because of an error
<a,b><failed attmpts at a or b or linker,c>

Note. The figure is created by the author based on Knoch's rating scale of coherence (2007b, p.180).

The results of the main analysis revealed that in the higher-scoring writing, superstructures and sequential progression tended to occur more frequently, while parallel progression occurred infrequently. On the other hand, there were unrelated progression and coherence breaks in lower or the lowest scored writing, which would agree with our intuition.

Some of the key coherence categories can be explained as follows.
As the rating scale indicates and Knoch (2007a) states in the conclusion section, superstructure and coherence break are categories that discriminate different levels of writing ability. In superstructure, which contributes to high scores, a linking device such as "For example" creates coherence effectively instead of topic progression. This category can be represented as <a, b><linking device, c, d>. In a coherence break, which occurs frequently in lower-level writing, an attempt at coherence fails because of an error. This category can be illustrated as <a, b><failed attempts at a or b or linker, c>. Furthermore, in addition to

superstructure, sequential progression, which is described as <a, b><b, c>, is observed frequently in higher scored writing, where "the comment of the previous sentence becomes the topic of the following sentence" (Knoch, 2007a, p. 115). In several previous studies, researchers have said that EFL learners are not good at proper sequential topic development (Belmonte & McCabe, 1998; Connor & Farmer, 1990; Kawanishi, 2019; Simpson, 2000; Wang, 2007).

In conclusion, Knoch's work is significant in that she developed descriptors based on empirically extracted coherence descriptors from actual writing samples based on the TSA approach to develop a rating scale specifically for coherence, which she said connotes "a fuzzy concept" (Knoch, 2007a, p.121; 2007b, p.97). Furthermore, this rating scale has been validated empirically. According to a many-facet Rasch measurement (MFRM) analysis using FACETS (Linacre, 2004), the raters rated more accurately with the TSA-based scale than the DELNA scale, and they used more band levels.

As described above, TSA truly "offers a productive approach to text analysis" (Schneider & Connor, 1990, p. 423). However, as limitations, TSA does not cover or explain all aspects of coherence (Knoch, 2007a; Schneider & Connor, 1990), just as one of the eight raters in Knoch's study commented in the questionnaire that the TSA-based rating scale was narrower than the DELNA coherence scale because it focused "only on topical structure and not on other aspects of coherence" (Knoch, 2007a, p. 121). Therefore, it is necessary to explore "how these types of topic progression relate to coherence through features such as elaboration, supporting details, and examples" (Shneider & Connor, 1990, p. 423), which may be achieved by capturing the passage from wider perspectives beyond local coherence.[9] Lastly, Knoch (2007a) states that “the (coherence break) category was created that accounts for features very specific to writers whose L1 is not English” (p. 114), namely, ESL learners. This is even more applicable to EFL learners in Japan, whose English proficiency levels should be much lower than Knoch’s ESL learners studying at an EMI university overseas. Finer analysis would be required for the category in question. Moreover, coherence breaks/anomalies must be explored in more detail in the context of the current study. This topic is reviewed in Section 2.4.4.

2.4.3.3 Development of Checklists for Diagnostic Assessment of L2 Writing. To conduct a diagnostic assessment of L2 writing, its components need to be closely explored and identified. It is also necessary to explore and conceptualize the common competences underlying these items, rather than their disparate existence.

Kim (2011) developed a diagnostic assessment scheme by analyzing 480 TOEFL iBT independent essays. She constructed a Q-matrix to identify to which subskill a descriptor should be assigned. As a result, she proposes the diagnostic assessment scheme known as the Empirically-Derived Descriptor-Based Diagnostic (EDD) checklist. It consists of 35 descriptors, which are grouped into five categories: content fulfillment (CON), organizational effectiveness (ORG),

[9] This section will not delve any further into the theory, but the concepts of the macrostructure of discourse and the schematic structures of discourse beyond the microstructure of discourse proposed by Kintsch and van Dijk (1978) are significant elements in terms of analyzing coherence in argumentative writing. In particular, the schematic structures of discourse, which is also called rhetorical genre analysis (Coulthard, 1994; Hatch, 1992), is significant.

grammatical knowledge (GRM), vocabulary use (VOC), and mechanics (MCH). Organization is almost always recognized as one of the writing subskills in rating scales. The criterion to which each descriptor belong is indicated. Although the results of the analysis indicate that some items have overlap, they are generally arranged in the order described above, starting with content fulfillment. Among them, 14 of the descriptors belong to ideational (i.e., content fulfillment, 1–8) and rhetorical (i.e., organizational effectiveness, 9–14) coherence. Table 2.9 is a list of 14 relevant items from the 35 items.

Table 2.9

A List of 14 Items Relevant to Ideational and Rhetorical Coherence

Content fulfillment according to Kim's feedback sheet	
1.	This essay answers the question.
2.	This essay is written clearly enough to be read without having to guess what the writer is trying to say.
3.	This essay is concisely written and contains few redundant ideas or linguistic expressions.
4.	This essay contains a clear thesis statement.
5.	The main arguments of this essay are strong.
6.	There are enough supporting ideas and examples in this essay.
7.	The supporting ideas and examples in this essay are appropriate and logical.
8.	The supporting ideas and examples in this essay are specific and detailed.
Organizational effectiveness according to Kim's feedback sheet	
9.	The ideas are organized into paragraphs and include an introduction, a body, and a conclusion.
10.	Each body paragraph has a clear topic sentence tied to supporting sentences.
11.	Each paragraph presents one distinct and unified idea.
12.	Each paragraph is connected to the rest of the essay.
13	Ideas are developed or expanded well throughout each paragraph.
14.	Transition devices are used effectively.

Note. Extracted From the EDD Checklist described by Kim (2011). Note that Kim's feedback sheet is called the Diagnostic EAP writing profile (p. 540).

The checklist described by Kim (2011) is very clear, concise, and highly commendable in that it provides important information as diagnostic feedback on what was done and not done in a way that is easily understood by learners. On the other hand, according to descriptor parameter estimates obtained from her cognitive diagnosis analysis, there is an overlap in as many as seven descriptors between content fulfillment and organizational effectiveness (items 2, 3, 4, 5, 7, 11, and 13). This overlap between content and organization suggests that the two components are somewhat inseparable and mutually dependent. When considered as a formative assessment, this issue would require a further step of analysis and instructional refinement based on this analysis to improve the learners' writing.

2.4.3.4 Taxonomy of Knowledge-Based Academic Writing Skills. As described previously, the information obtained from the exploration of the components of writing skills and their taxonomy is an essential research topic for writing assessment and education. Grabe and Kaplan (1996) proposed a model of text construction and parts of a taxonomy of academic writing skills, which is classified on the knowledge-based writing skill types. They collected the information through an ethnography of writing and categorized it into a taxonomy of writing skills and contexts. Although it is not in the form of a rating scale, their study is valuable as it explores the components of writing skills and provides the basis for creating a rating scale.

Table 2.10 presents the taxonomy generated by Grabe and Kaplan (1996), which is transcribed and modified by the author with a focus on discourse knowledge, which is assumed to involve coherence-related issues. This taxonomy shows that there are many different aspects and variables in writing skills (six skills with as many as 20 descriptions) including both local and global coherence. However, it does not suggest any hierarchical structure, which means “it is not sufficient to be used as a basis for the development of rating scale criteria by themselves” (Knoch, 2011, p. 86). Moreover, the nine statements (labeled a through i) under discourse knowledge seem to be mixed, varying from cohesion to rhetorical and ideational elements without clear distinction. Furthermore, it should be noted that this taxonomy concerns L1 and L2 English writing. It appears to be a more conceptual rather than practical classification, in contrast to the descriptors provided by Kim (2011).

Table 2.10

Taxonomy of Writing Skills and Contexts Featuring Discourse Knowledge

1		**The writer’s circumstances intentions, goals, attributions and attitudes**
2		**Linguistic knowledge** (a–f)
3		**Discourse knowledge**
	a.	Knowledge of intra-sentential and inter-sentential marking and devices (cohesion, syntactic parallelism)
	b.	Knowledge of informational structuring (topic/comment, given/new, theme/rheme, adjacency pairs)
	c.	Knowledge of semantic relations across clauses
	d.	Knowledge to recognize main topics
	e.	Knowledge of genre structure and genre constraints
	f.	Knowledge of organizing schemes (top-level discourse structure)
	g.	Knowledge of inferencing (bridging, elaborating)
	h.	Awareness of differences in features of discourse structuring across language and culture
	i.	Awareness of different proficiency levels of discourse skills in different languages
4		**Sociolinguistic knowledge** (a–e)
5		**Further audience considerations**
6		**Knowledge of the world**

Note. Adapted from Grabe and Kaplan (1996).

2.4.3.5 Defining L2 Writing Components Through Raters' Decision-Making Behaviors. The attempt to extract and categorize the evaluation criteria of L2 learners' written products by exploring the decision-making process of raters in writing evaluation is an effective approach to identify conceptual definitions of writing skills. From this viewpoint, previous research has been conducted to investigate raters' behaviors while evaluating writing with existing rating scales (e.g., Cumming, 1990; Lumley, 2002; Milanovic et al., 1996). Milanovic et al. (1996) conducted their survey on raters' behaviors while using holistic rating scales for writing evaluation, whereas Cumming (1990) and Lumley (2002) investigated raters by using multiple-trait rating scales. Lumley (2002) reported that raters struggled to bridge the gap between existing rating scales and their own intuitive evaluation norms. Unlike the above studies, where raters use existing rating scales, Cumming et al. (2001, 2002) conducted their investigation of raters' behaviors while evaluating L2 writing with no specific rating guidelines. This section reviews their study. This review is partially adapted from the author's paper (Matsumura & Takagi, 2022, p. 47).

Cumming et al. (2001, 2002) conducted a series of studies to develop a descriptive framework on the decision-making processes of raters as part of the development process of TOEFL 2000, including interrelated projects to design a new TOEFL. This study consisted of three sub-studies. The first included the development of a preliminary descriptive framework based on the think-aloud protocols by 10 experienced ESL/EFL assessors. These raters evaluated 60 TOEFL essays, where their verbal protocols were collected while rating. In the subsequent second study, the framework was reviewed for refinement by supplemental think-aloud data provided by another seven experienced raters. The third study aimed to refine the framework. The authors developed and refined the descriptive framework through multiple stages of experimentation.

Table 2.11 shows the result with the raters' decision-making behaviors categorized into three foci: self-monitoring focus, rhetorical and ideational focus, and language focus. The authors further classified them into two strategies, interpretation strategies and judgment strategies, with multiple decision-making behavioral elements. The framework includes 35 distinct and independent elements.

Table 2.11

Descriptive Framework of Decision-Making Behaviors While Rating TOEFL Writing Tasks

Self-monitoring focus	**Rhetorical and ideational focus**	**Language focus**
(descriptions omitted)	***Interpretation strategies***	
	(omitted)	(omitted)
	Judgment strategies	
	Assess reasoning, logic, or topic development	Assess quantity of total written production
	Assess task completion or relevance	Assess comprehensibility and fluency
	Assess coherence and identify redundancies	Consider frequency and gravity of errors
	Assess interest, originality, or creativity	Consider lexis
	Assess text organization, style, register, discourse functions, or genre	Consider syntax or morphology
	Consider use and understanding of source material	Consider spelling or punctuation
	Rate ideas or rhetoric	Rate language overall

Note. The author has omitted descriptions that have little relevance to this study (self-monitoring focus in the left column and interpretation strategies in the upper right row). Based on Barkaoui (2007, pp. 104–105) and Cumming et al. (2001, p. 53; 2002, p. 88).

As a study implemented based on the framework of Cumming et al. (2001, 2002), Barkaoui (2007) investigated how different types of rubrics, holistic and analytic, affect rater behavior while assessing L2 argumentative writing products. The authors concluded that the holistic scale resulted in higher inter-rater agreement in terms of decision-making. Aside from the results of his research on the differences in the rating scales, Barkaoui's (2007) empirical study, which is a replication of the study by Cumming et al. (2001, 2002), showed that the descriptors in the framework functioned as a measure of classification when raters evaluated L2 argumentative writing.

2.4.4 Identifying Coherence Breaks/Anomalies in Formative Assessment

2.4.4.1 Coherence Breaks/Anomalies in L2 Writing. This section provides a review of studies on the analysis of coherence breaks/anomalies, mainly for EFL learners. Coherence breaks are also discussed in the context of Knoch's (2007a, 2007b) TSA (Section 2.4.3.2). As she explains, coherence breaks, when "attempt at coherence fails because of error" (Knoch, 2007b, p. 124), are created mainly for L2 learners. Needless to say, those whose first language is not English are more likely to have coherence breaks in their L2 writing. In her research, coherence breaks are grouped together into one category and packaged as a phenomenon found in writing products earning the lowest rating. This may be a deserved treatment, in a sense, given that English language learners with an inherently low overall language ability are not considered to have coherence ability, as stated in the CEFR descriptors reviewed in Section 2.4.2. However, a diagnostic assessment of Japanese learners of English for the purpose of providing feedback for remediation would require more detailed analysis and diagnosis of these coherence breaks. This is not difficult to understand when reminded of the elaborate intermediate and lower intermediate descriptors on the large-scale standardized tests that are unique to Japan. Therefore, it is necessary to further review previous studies on coherence breaks.

Moreover, a continuing major concern of English writing researchers and instructors has been the factor(s) that distinguish the higher and lower scored writing, including the quality of organization. Error approach or error analysis, that is, examining errors that deteriorate the quality of writing, is one of the approaches to address this issue (Corder, 1967; Maimon & Nodine, 1978). However, the sources of errors are often complex and diverse, so what they can do with certainty is sometimes limited to explain the score in relation to the frequency of error occurrence (Witte & Faigley, 1981). It is not an easy task to measure the degree of severity of errors. However, this endeavor is required to evaluate organization of discourse because the quality of organization is not a matter of the right or wrong of the alternative, but often the degree of appropriateness. Moreover, there is no one way to correct inappropriate parts in terms of coherence as remediation, that is, it is possible to change or revise them in various ways. Alderson (2005, p. 260) summarizes these persistent problems with error analysis into the following five points: difficult in identifying (a) what the learner was trying to say, (b) what the source of errors was, (c) whether the error was persistent of intermittent, (d) under what performance conditions the error occurred, and (e) why the error occurred. In other words, error analysis requires consideration of many perspectives. Considering coherence breaks/anomalies, issues on error analysis is mentioned here before introducing some empirical studies. As Spillner (1991) describes in his bibliography of error analysis, the failure of error analysis to live up to its promise means that it has been abandoned. Alderson (2005) believes, however, that its feasibility enables the creation of more comprehensive and detailed longitudinal and cross-sectional corpora of learners of various proficiency levels. He suggests that the existing corpora had insufficient detail.

In light of the above, one way to overcome the previously mentioned difficulties in organization assessment may be to examine the learners' writing closely and to determine where the problems lie and investigate each characteristic to correlate them with the scores. Elaboration of work to detect organizational problems that break coherence of discourse can be one of the key issues for communicative writing instruction. Accordingly, some researchers have attempted to identify and categorize discourse anomaly on texts produced by EFL learners.

Wikborg (1990) led one of the key studies in identifying coherence breaks of essays written by EFL students. She defines "coherence breaks" as "what happens when the reader loses the thread of the argument while in the process of reading a text attentively" (p. 133). She introduces two types of coherence breaks: topic-structuring problems and cohesion problems with a total of eleven subcategories. In her study with 114 essays written by Swedish EFL students, the five most frequent types of coherence breaks accounted for 82% of the total 801 instances: (a) uncertain inference ties, (b) misleading paragraph division, (c) missing or misleading sentence connection, (d) unjustified change of / drift in topic, and (e) unspecified topic. While her study is very informative, it would be preferable if the relationship between these anomaly breaks could be shown a little more clearly.

2.4.4.2 Identifying Coherence Breaks/Anomalies in L2 Writing with Annotation Tools. Skoufaki (2009) investigated rhetorical anomaly among texts produced by EFL students. She detected coherence errors in 45 paragraphs written by Chinese EFL students. What is notable about this study is that she utilized rhetorical structure theory (Mann & Thompson, 1988) for her investigation. Moreover, she attempted to investigate the extent to which the errors detected by using rhetorical structure theory (RST) analysis matched those located by criterion (Burstein et al., 2004), a well-known automated writing evaluation (AWE) software.

Ahmadi and Parhizgar (2017) also employed RST to detect coherence anomalies. They examined Iranian EFL learners' writing samples to find eight different types of coherence errors. The authors concluded that irrelevance and change of a topic are the most frequent type of coherence breaks and that these can be partly ascribed to the EFL learners' essays in an inductive order. Other researchers have also used RST to detect coherence breaks in English writing by L1 English students (e.g., Candlin et al., 1998; O'Brien, 1995), which suggests that the coherence issue is not necessarily limited to L2.

Yamashita (2019) led one of the most recent studies of detecting coherence anomaly with the RST. She used RST to detect organizational anomalies in Japanese university students' writing samples and categorized them into eight groups. She found that irrelevant ideas and sudden topic shift occurred most frequently. Based on these results, she concluded that there are three causes of these anomalous sentences in Japanese EFL student writers: (1) missing or non-functional topic sentence due to inductive logical development, (2) redundancy often found in the Japanese text, and (3) dependence on readers' inference often found in the Japanese text.

Lastly, Kawase (2020) reported on his ongoing research where he has his students use RST to analyze their own writing as EFL writing class activities at a Japanese university. Unfortunately, this study is a practice report not based on a research design that can be validated, so the effectiveness of the actual activities with RST still needs to be investigated.

As seen above, many writing researchers have used RST as an annotation tool. The use of annotation tools in language studies is discussed in Section 2.5.1.3, but we begin with a review of previous research on graphic displays in learning.

2.5 Text Annotation Tools and Graphic Displays as Structural Representation

As mentioned in the previous section, the detection of coherence breaks and comparing and contrasting them with different groups of learners would be helpful in L2 writing research and would provide learners with many useful suggestions. At the same time, a more in-depth analysis of anomalies may be desired. Discourse annotation, which intends to create a structured representation from the text, would be a powerful means to achieve the objectives to analyze and understand the structure of the passage or detect and locate the problems in terms of organization. An annotation tool describes the interrelationship of discourse units (sentences, clauses, or other kinds of segments) and presents their role in the overall discourse (Mann & Thompson, 1988; Wolf & Gibson, 2005). Before discussing how annotation tools are actually used in the evaluation and analysis of writing, the research on the value that graphical displays have in learning is reviewed.

2.5.1 Graphic Displays in Language Learning

2.5.1.1 Theoretical Background of Graphic Displays in Learning. Graphic displays, which are also called visual displays, graphics, and graphical representations interchangeably, are characterized as displays "that represent objects, concepts, and their relations using symbols and their spatial arrangement" (Vekiri, 2002, p. 262). Graphics are monosemic, which means they have a single meaning, so they are distinguished from other symbolic systems such as pictorial representations (Bertin, 1983).

In her review of graphical displays in learning, Vekiri (2002) summarizes three theories based on information processing approach: the visual argument hypothesis, the conjoint retention hypothesis, and dual coding theory (p. 263). While the visual argument hypothesis presupposes that graphical displays are more effective in perceptual organization "in communicating information about data relations, trends, and patterns" than text, conjoint retention and dual coding focus on effectiveness on retrieval of information from memory. Among them, the visual argument hypothesis explains clearly how graphics effectively help learners. This hypothesis is most relevant to language learning. According to the visual argument hypothesis, "graphical representations are effective because, owing to their visuospatial properties, their processing requires fewer cognitive transformations than does text processing and does not exceed the limitations of working memory" (Vekiri, 2002, p. 281). Moreover, graphical representation makes it easier for users to perceive or draw inferences about individual elements and their relations compared with text comments (Robinson & Kiewra, 1995; Winn et al., 1991). The experimental design that is most relevant to one of the RQs of the present study based on the visual argument theory is that used by Winn et al. (1991) (Table 2.12). Among their sample of graduate students, they compared the time required to solve a problem between an experimental group presented with a tree diagram and a control group presented with text. The experimental group took less time to solve kinship problems. The limitation was that the effect disappeared when students were not familiar with the conventions and terms of the diagrams. The authors inferred that graphic presentation may not be a good fit for all students, and that there seem to be students that find it a good fit for some and not for others.

Table 2.12

Summary of a Visual Argument Study Relevant to the Present Study

Study	Winn et al. (1991)
Display	Tree diagram
Participants	Graduate students
Learning outcomes	Response latencies (time needed to solve each problem)
Instructional conditions	Students were asked to solve kinship problems using either tree diagrams or lists of sentences.
Findings	Students took less time when using diagrams.

Note. Adapted from Vekiri (2002) and Winn et al. (1991).

The visual argument hypothesis that presupposes "owing to their visuospatial properties, their processing requires fewer cognitive transformations than does text processing" (Vekiri, 2002, p.281) and the experiment result regarding the effectiveness of graphical display on time on task (Winn et al., 1991) led to one of the RQs in the present study related to time on task.

2.5.1.2 Graphic Organizer as a Scaffolding for Reading Comprehension. A well-known example of the use of graphic displays in language learning is a graphic organizer. It is defined as a graphic depiction of the relationships between concepts in a text (Kools et al., 2006) or as tools to help readers appreciate the content of the text by analyzing the whole structure of the passage (Grabe, 2004; Jang & Grabe, 2007). A graphic organizer is also thought to be effective in assisting and facilitating comprehension of the content when reading texts, and this is used regardless of whether the text is L1 or L2. There have been attempts to visualize the passage for better understanding of the text in reading instructions.

As an illustrative example of a graphic display used for L2 language learning, Ishii (2006) conducted an experimental study with 40 Japanese-as-foreign-language (JFL) learners divided into the graphic presentation group, the incomplete graphic completion group, and the control group. The graphic display employed in this study was two types of charts prepared to illustrate the historical descriptions in the form of flow charts with boxes and arrows. The participants were asked to reproduce the written text in their first language after reading a history text of approximately 1,800 characters in Japanese. The evaluation was based on two indices: the hierarchy of the text structure and the overall understanding of the text from the beginning to the end. The results showed that the graphic presentation group reproduced significantly more text than the control group, and the incomplete graphic completion group fell between the other two. The results suggest that the presentation of graphics could help L2 learners to select and structure the important ideas. In conclusion, Ishii (2006) discussed that L2 learners have limited cognitive resources, so it is not always easy for them to consistently understand and remember text content. Therefore, graphic displays seem to reduce the burden of language processing, enhance text comprehension, and facilitate information integration.

Jiang (2012) used a graphic organizer for EFL learning a well-known reading assistance strategy. After one semester of instructing 340 Chinese EFL learners on a discourse structure graphic organizer, the scores increased significantly on an independent reading comprehension test and on the TOEFL reading test immediately after instruction. However, the retention effect was lost on the delayed TOEFL test administered to the same population, although the retention effect was observed on their original reading test in the classroom.

The aforementioned studies suggest that graphics help learners to understand the content and the discourse structure of L2 text. However, the retention of the effect may be limited, and more comprehensive research analysis on the scope and conditions of the effect is needed.

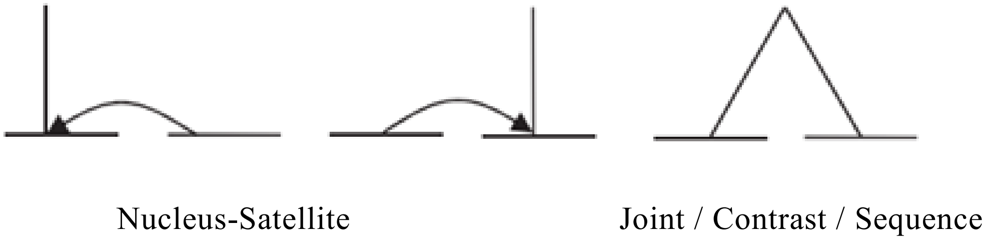
2.5.1.3 Graphic Displays with Annotation Tools for Analysis of L2 Learners' Written Products. The focus of the present study is writing, not reading, but the revision work in process writing requires the ability to read one's own draft critically. In other words, revising a draft presupposes the ability to read it. The rewriting process can be decomposed into two steps: rereading and revising. Therefore, graphic displays, which would assist reading comprehension, could also be helpful in rewriting the text.

Researchers have noted that the effectiveness of graphic displays for L2 learners has not been sufficiently investigated in reading research (Jiang & Grabe, 2007). However, as presented in Section 2.4.4.2, several L2 writing studies have utilized graphic displays to detect and categorize coherence errors in EFL writing (e.g., Ahmadi & Parhizgar, 2017; Kawase, 2020; Skoufaki, 2009; Yamashita, 2019). Among the annotation schemes, rhetorical structure theory (RST) (Mann & Thompson, 1988) has been one of the most prevalent.

RST is a theory of relational structure that expresses the organization of coherent contiguous text (Mann et al., 1992). It was generated based on the empirical study with more than 400 English texts from academic writing. The basic RST schema consists of a nucleus–satellite combinations, where the satellite supports the nucleus (Figure 2.3). Moreover, the hierarchical structure consists of a series of nuclear–satellite combinations. The satellite, which is a source text, supports the target text unit called a nucleus, whose relation is hierarchical. A satellite text unit is linked to a nucleus text unit via a specific relationship (e.g., condition, means, preparation, etc.). In addition, RST text units are not sentences; rather, they are usually at the level of phrases or clauses. The multilayered structure is presented in Figure 2.4. The labels on the arc indicate the relationship between satellite and nucleus.

Figure 2.3

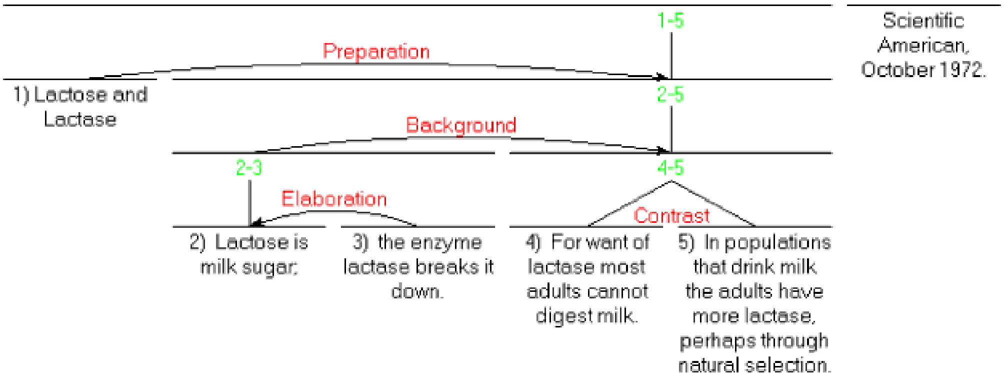
Nucleus–Satellite Schema of RST



Note. Schemas in RST presented in Kawase (2020, p.38).

Figure 2.4

Multilayered Hierarchical Diagram Obtained by RST


1-5
Preparation
Scientific American, October 1972.
1) Lactose and Lactase
2-5
Background
2-3
4-5
Elaboration
Contrast
2) Lactose is milk sugar;
3) the enzyme lactase breaks it down.
4) For want of lactase most adults cannot digest milk.
5) In populations that drink milk the adults have more lactase, perhaps through natural selection.

Note. Lactose example presented in Mann (2003)

Thus, dividing EFL learners' writing into text units and showing the connections through tagging with RST would help to identify the relationship between the text units and provide an overview of the text's structure. Annotation tools represent a powerful approach to perform this task. However, there are several issues, particularly the textual unit classification of RST. One issue is that there are as many as 23 options (Taboada & Mann, 2006) in RST taxonomy for labeling the connection, so there could be inconsistencies or disagreements in the labeling of relationships both within and across annotators. Moreover, the definitions of those options may occasionally require knowledge of English linguistics, which may make it difficult for non-researcher classroom instructors to make decisions on labeling. In addition, the RST classification and its diagram may be too complex to allow learners to clarify what is wrong with the sentence connections. A simpler annotation tool would be desirable in educational settings.

2.5.1.4 Differences Between Graphic Organizers in Reading Research and Annotation Diagrams in Writing Research. There have been attempts to visualize a passage to better understand the text in the language learning context. In this sense, diagrams generated by discourse annotation tools in L2 writing research and graphic organizers in L1/L2 reading research in the language learning context may be in the same category. However, what differentiates the two are the purpose and the grain size of the text unit.

One difference between reading and writing research is the user of the graphic displays. Whereas the primary users of the graphic displays in reading research is the learner, the user in writing research is often the researcher, instructor, or annotator. While reading researchers encourage learners to comprehend the text by presenting or having them create a graphic display, writing researchers use a graphic display to evaluate and analyze learners' written products. Moreover, writing researchers require some expertise and training regarding annotation tools because each tool has its own functions and tagging rules. In contrast, reading researchers can adapt graphic organizers according to the purpose, as long as the researcher has certain theoretical background knowledge of the framework.

Another difference between reading and writing research is the granularity of text segmentation. While the granularity of text units in graphic organizers for reading comprehension is usually much larger and decided freely by the instructors depending on their reading purposes, discourse annotation tools focus on examining how sentences form a flow of meaning (Grosz & Sidner, 1986), and require finer granularity. While graphic organizers and annotation tools are the scheme for reading comprehension. The relationship between ideas in text is formatted into a diagram by decomposing the text in mind mapping (Putra et al., 2021), works the other way around as it is used for text generation. It helps writers build up or organize the structure of the text.

2.5.1.5 Types of Annotation Tools Used to Generate Annotation Diagrams. Discourse annotation aims to create a structured representation of the text to explain how the discourse units relate to each other (e.g., Mann & Thompson, 1988; Wolf & Gibson, 2005). In the annotation process, the user (annotator) annotates or labels data by tagging with relevant metadata. Regardless of the annotation tool, the process of text annotation basically follows three main steps: text segmentation, tagging, and connecting text units. Putra et al. (2021) describes a range of annotation tools developed by researchers in the natural language processing (NLP) community. There have been various annotation tools developed specifically aiming at global discourse annotation, each of which is different in many functions and usability with different purposes, target text genres or target users. In other cases, the use of the language is not for learning purposes, but rather for discourse analysis of difficult and complex legal documents and the like. The differences in annotation tool development and functionality are technical and beyond the scope of this study and will not be discussed further here. However, a comparison of RST generated by Mann and Thompson (1988), which is commonly used by language researchers, and TIARA developed by Putra et al. (2020, 2021) and used in this study for text analysis, along with some other annotation tools,[10] are presented in Table 2.13. Because this study does not aim to

[10] Tool A: Grapat (Sonntag & Stede, 2014); https://github.com/discourse-lab/GraPat
Tool B: DiGAT (Kirschner et al., 2015); https://github.com/UKPLab/
Tool C: OVA (Janier et al., 2014); https://www.arg-tech.org/index.php/ova/
Tool D: TreeAnno (De Kuthy et al., 2018). https://gitbub.com/nilsreiter/treeanno
RST: https://www.wagsoft.com/RSTTool/
TIARA: https://github.com/wiragotama/TIARA-annotationTool

study the details of the annotation tools, they are presented anonymously in the table as examples, but for reference, each specific tool is listed in the footnotes.

TIARA is a strong annotation tool for relation-focused discourse annotation with a tree structure. As can be seen in Table 2.13, there is no gold-standard discourse annotation tool. First, RST seems to have preferable features, but it has drawbacks in terms of usability and simplicity of relation options as is discussed in the previous section. Among the annotation tools of graph structure, Tool C seems to be TIARA's strongest competitor. It offers as many features as TIARA does, with the exception of the discourse unit reordering feature and discourse unit categorization. In particular, discourse unit categorization (the reordering function) is used neither for assessment nor for feedback, but, if preferred, it could be used in the future instruction. This feature would be very useful if students themselves were to use the tool to make revisions in the future.

Table 2.13

Comparison of the Features of RST, TIARA, and Other Discourse Annotation Tools

Feature	Tool A	Tool B	Tool C	Tool D	RST Tool	TIARA
1. Discourse structure	Graph	Graph	Graph	Tree	Tree	Tree
2. Segmentation			√		√	
3.AC and no-AC categoraization	√	√	√			√
4. Discourse uint categorization	√					√
5. Linking	√	√	√	√	√	√
6. Link labelling/polarity	√	√	√		√	√
7.Structure visualization	√	√	√	√	√	√
8. Annotation scheme cusomization			√		√	√
9. Discourse unit reordering						√
10. Text editing			√		√	√

Note. Tool A, GraPat; Tool B, DiGat; Tool C, OVA; Tool D, TreeAnno
AC, argumentative component. Adapted from Putra et al (2021).

2.6 Relevant Rating Scale for Diagnostic Assessment

The importance of the rating scale is an undisputable fact. Language assessment is discussed based on the premise of the validity of the rating scale, and various statistical methods for validating the rating scale have become a major interest in the language assessment field. Furthermore, the purpose of the assessment, the usability of the rater, and the economics of time and effort must also be considered when selecting the rating scale to be used in classroom assessment studies.

Regarding the rating scale design process, Weigle (2002, pp. 122–124) suggests the following issues be considered: (a) Who is going to use the rating scale? (b) What aspects are most important and how will they be divided up? In other words, what are the criteria based on? (c) How many points, or scoring levels, will be used? (d) How will scores be reported? Even if developing a rating scale is not the purpose of this study, the most appropriate rating scale that satisfies those conditions can be found by answering these questions as a viewpoint for selecting a rating scale. Moreover, it would be helpful to consider these questions when it is necessary to make slight modifications to the existing rating scale to fit the research objectives or subjects.

2.6.1 Who is the Rating Scale for?

There are three types of rating scales depending on who is going to use them (Alderson, 1991, p. 72). The first subcategory is a user-oriented scale, which gives information on what the score indicates in terms of the test takers' ability to perform the targeted language skill such as can-do statements and checklists. The second subcategory is an assessor-oriented scale, which guides raters through the rating process to help them decide which level a test-taker should be labeled based on the descriptors of each level. Lastly, constructor-oriented scales are developed for test constructors to set appropriate tasks for corresponding levels of the test needed. Although these scale types should be used according to the purpose of research, as Knoch (2007a) indicates—by referring to North (2003)—that assessor-oriented scales should be used for second/foreign language performance assessment among the three types of scales in general. This is even more true with classroom-based assessment because "such scales provide guidance not only on how to rate the performance, but also on what kinds of tasks to present to candidates in order to elicit performance that can be rated" (Bachman & Palmer, 2010, p. 344). However, in diagnostic assessment, the most important perspective to be considered is the learner's viewpoint, and in that sense, the user-oriented scale perspective should be given priority as well as instructors or raters. Students are the most important stakeholder in assessment, and it is the learners who are most affected by the consequences.

2.6.2 What Aspects of Writing Are Most Important, and How Will They Be Divided Up?

Many researchers have discussed the advantages and disadvantages of various scales (e.g., Bachman & Palmer, 1996; Cohen, 1994; Fulcher, 2003; Grabe & Kaplan, 1996; Hyland, 2003; Kroll, 1998). There are two main types of rating scales in terms of approach of scoring based on Weigle (2002): holistic and analytic. Each approach has its advantages and disadvantages, and writing test constructors decide on the approach that matches their purpose. Holistic scales remain popular due to their cost effectiveness, particularly for proficiency testing in high-stakes contexts (Weigle, 2002). The simplicity of a single score outcome is convenient to explain the result to the test takers for ranking based on an overall assessment. Still, it is possible for those large-scale standardized tests to provide the test takers with diagnostic feedback based on the descriptors in the rating scale.

According to Weigle (2002), the quality of an analytic scale can be evaluated in terms of reliability, construct validity, practicality, impact, and authenticity. As advantages, it shows higher reliability than a holistic scale. For this reason, an analytic scale is more appropriate for L2 writers as different aspects of writing ability develop at different rates, and it allows providing information required for placement and/or instruction. However, there are also some disadvantages. For example, it is time-consuming and expensive to score when using an analytic scale, and raters may read students' writing holistically and adjust analytic scores to match their holistic impression.

Considering the above issues, analytic scoring is the determine what types of rating scale are desired for formative diagnostic assessment. Analytic scoring is time-consuming and may have some problems in terms of authenticity, but useful diagnostic information is more likely to be obtained. As discussed in Section 2.3, a desirable rating scale for diagnostic assessment/feedback is the one which can provide detailed information about the target language.

2.6.3 What Are the Criteria Based on?

The rating scale development can be divided into two main groups based on the design methods: theory- or measurement-driven method, and empirical- or performance-based method. In addition to these two, there is another method called intuition-based method. They are illustrated in Table 2.14.

In a theory-based method, a rubric is created directly from the experience and knowledge of experts. In contrast, in an empirical-based method, a rubric is created empirically based on learner performance, which is called a performance-based method (Fulcher et al., 2011). In a theory-based method, descriptors are developed *a priori*, so they may sometimes be ambiguous and discrepant with the performance of the target population (Fulcher, 1996; Turner & Upshur, 2002; Upshur & Turner, 1995). The mismatch between the descriptors and the rating scale in a small population such as students in a classroom might be a critical issue when a rubric is used without modification. As mentioned in Chapter 1, the combination a theory- and performance-based method is applied in this research, which can better reflect the reality of student writing as well as the reliability of the rating scale supported by conceptual background. Therefore, the existing theory-based rating scale has been modified to provide the students with formative diagnostic feedback through classroom-based assessment.

Table 2.14

Rating Scale Categories Depending on the Development Method

Types of development	Types of methods/models	Approach	Examples
Intuition based (intuitive meth	(a) Expert judgement (b) Committee method (c) Experiential design	(a) Developed by expert teachers/test developers based on existing rating scales, syllabus, needs analysis (b) Developed by a small group of experts (c) Evolved and reined of the existing rating scale	FSI family of rating scales: ILR, ACTFL, ASLPR, and IELTS
Theory based (measurement based)	(a) The four-skills model (b) Models of communicative competence (c) Theories of writing (d) Models of decision-making by expert judges/raters	(a) Developed based on simple, generic, and familiar conceptual framework such elements as mechanics, vocabulary, grammar, and organization (b) Developed based on the assessment criteria on a model of language ability (c) Developed based on research on writing as a product or process (d) Developed based on a protocol analysis of raters regarding their decision-making behavior	Madsen (1983); Jacobs et al. (1981) Bachman (1990); Bachman and Palmer (1996) Grabe and Kaplan (1996) Sakyi (2000); Cumming et al., (2001, 2002)
Empirical based (performance based)	(a) Data-based or data-driven (b) Empirically derived, boundary-choice, boundary definition (EBB) (c) Scaling descriptors	(a) Developed based on analysis of performance on tasks, and the description of key features of performance (b) Developed by asking expert judges to divide samples into better and poorer performance (c) Developed empirically by following four distinct phases consisting of activities such as collecting, grouping, analyzing qualitatively, and establishing cut-off points with samples	Fulcher (1993, 1996, 2011) Upshur and Turner (1995, 1996); Turner and Upshur (2002); North (1995, 2000); North and Shneider (1998)

Note. ACFTL, American Council on the Teaching of Foreign Languages; ASLPR, Australian Second Language Proficiency Ratings, FSI, Foreign Service Institute; IELTS, International English Language Testing System; ILR, Interagency Language Roundtable Summarized by the author based on Fulcher (2003) and Knoch (2007b).

2.6.4 What Will the Descriptors Look Like and How Many Scoring Levels Will Be Used?

According to Knoch (2007b), there are three issues to be considered regarding the appearance and scoring band levels in the development of rating scales: the number of bands in a scale, distinctions between levels, and descriptor formulation style. Each of these issues is discussed briefly.

There are two issues that should be considered when deciding the number of band levels in a rating scale; one is a matter of relationship between reliability in ratings and decision power or tension of raters. Miller (1956) suggests that human processing capacity in general in differentiating samples into levels is around seven (plus or minus two) in rating scales, categories for absolute judgment, and objectives in the span of attention or of immediate memory (p. 96). Myford (2002) agrees with this by suggesting that the reliability was high for scales raging from five to nine. As the second point, what is more important to consider how many band levels are appropriate to evaluate the targeted language skill. This number depends on the assessment purpose/context, and it differs from scale to scale. First, it would be necessary to reconfirm whether the level matches the target students and whether the test is set up to discriminate between that many levels. Then, remediation as a statistical method can be considered. Interactional adjustment or revision of the number of the levels after conducting the validation study with Rasch analysis may be one solution.

2.6.5 How Will Scores Be Reported?

For formative diagnostic assessment, just reporting a score is never sufficient, whether it is a single combined score or multiple scores of sub-categories. The test takers need to understand what the score means, and the test developers are accountable for ensuring the test takers understand. Scores should be accompanied by other diagnostic information. One of the options of diagnostic score report is a form of checklist. Kim' s (2011) checklist provides test takers with information of mastered and non-mastered items. She applied a psychometric CDA model, whose theoretical background is different from this study. However, her results can be fully utilized especially for items other than content and organization, where the difference between mastered and non-mastered items is clear and distinct. Kim's (2011) diagnostic feedback sheet based on the Empirically-Derived Descriptor-Based Diagnostic (EDD) checklist consists of 35 concrete and fine-grained descriptors of five language skill criteria: content fulfillment (CON), organizational effectiveness (ORG), grammatical knowledge (GRM), vocabulary use (VOC), and mechanics (MCH).

Another rating scale to be mentioned is the ESL CP (Jacobs et al., 1981). This is a commonly used analytic writing scale, and it is also used as the baseline to develop a tentative rating scale in this research. Each component of the scale is weighted according to its relative importance to the overall performance determined by the testing program. The scale consists of five criteria, namely content, organization, vocabulary, language use, and mechanics, with weighted scores of 30, 20, 20, 25, and 5, respectively. The total weight for each component is further broken down into numerical ranges that correspond to four master levels: excellent to very good, good to average, fair to poor, and very poor (p. 91). A notable advantage of this scale is that it provides key-word descriptors that help raters assess students' performance with ease. The detailed descriptors also contribute to distinguishing and characterizing each level.

Despite the advantages of the aforementioned scale, there are two main issues. Some of them are shared with other scales and others of which are particular to Jacobs' scale. One is about descriptors in relative expressions with adjectives or adverbs. For example, the phrases like "fluent expression" or "clearly stated" are used as descriptors in the organization criterion. Because the rater has to score multiple criteria in order, some researchers are concerned about a halo effect, which is often pointed out with analytic scoring. The other issue that applies particularly to Jacobs' scale is the difficulty in allotting a certain score in each score band. For example, it is not easily explained what differentiates between a score of 23 and 24. Moreover, the pilot study suggested that some descriptors essentially overlap with each other, which sometimes leads to obscure the cause of performance problems.

Each type of rating scale has its strengths and weaknesses, and there is no one-size-fits-all solution. It is a matter of trade off and prioritization. Moreover, test context is another key issue to be considered, for example, whether it is high- or low stakes assessment. In terms of relevance to this study, the rating scale from Jacobs et al. (1981) would be appropriate as the baseline for analyzing performances as well as for providing diagnostic feedback. However, according to the pilot study conducted by the author, the four band levels of Jacobs' rating scale were not necessarily used effectively. Thus, levels have been modified to make them relevant for assessing the performance of the targeted students.

2.7 Study Hypotheses and RQs

This section presents hypotheses based on the literature review as they relate to issues raised in the previous studies and to the RQs to be addressed in this study.

As described in Section 1.1, difficulties in assessing, teaching, and learning coherence in L2 writing skills have been noted (e.g., Cerniglia et al., 1990; Lee 2002) with factors including ambiguity of definitions and components extending across sentences and across entire passages. This also affects the provision of effective feedback to learners, although feedback is essential to students' writing development (Biber et al., 2011; Ferris, 2003; Hyland, 2003; Montgomery & Baker, 2007).

The literature review in this chapter included six main categories: current issues of EFL writing in Japan, writing feedback, diagnostic language assessment on EFL writing, coherence in writing, text annotation tools and graphic display as structural representation, and relevant rating scales for diagnostic assessment.

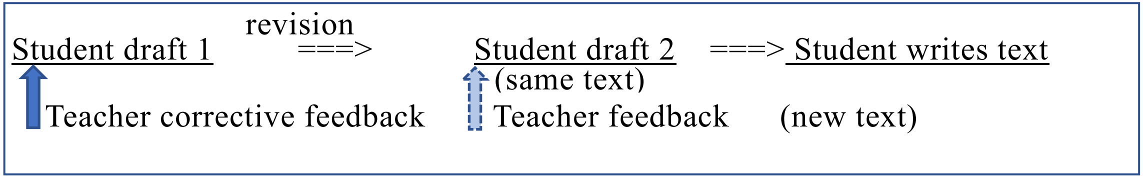
First, the current issues of EFL writing in Japan were addressed in Section 2.1. While the problem of Japanese people's ability to communicate in the world has been pointed out, how EFL writing is taught and learned in Japan has attracted attention, including the introduction of a new subject called Logic and Expression to the high school curriculum in the reformed MEXT Course of Study and an opinion writing section in the Eiken test, the large-scale standardized test that has the highest number of examinees in Japan. However, there are no consistent guidelines for teaching writing in universities as there are in high school, and surveys have shown that most university students are the least confident in their writing skills among the four skills. Therefore, considering these factors, this study focuses on writing skills, especially in coherent English, which requires logical development. The hypothesis is that argumentative writing instruction in a university setting is effective, and teacher feedback based on formative diagnostic assessment is helpful for learners who lack confidence in their ability.

Section 2.2 contains a review of writing feedback to explore what kind of feedback would be desirable. In corrective feedback, which assumes correction of process writing, there have been controversies over the effectiveness of grammar correction, as symbolized by the Ferris–Truscott controversy. There are several types of assessment design incorporating feedback in terms of focus or timing: form-focused versus content-focused feedback, and L2 writing, SLA design, or blended design. Among the various arguments, Truscott (1996) points out three important points to keep in mind when conducting L2 writing assessments: developmental sequence in the acquisition of grammar, consistency in the teacher's feedback, and avoidance of excessive corrections, which would hinder the learner's writing fluency and willingness to produce text. Based on the previous studies, the present study does not focus on one specific area, but rather addresses multiple areas in a comprehensive manner. Moreover, this study provides extensive feedback in the rhetorical areas that have been overlooked.

Considering the timing to provide feedback, it would be inappropriate to give students no feedback on the second draft after having them do an assignment as in the possible blended design Ferris (2010) proposes, which are introduced in Fig. 2.1 in the present study. Moreover, it would be inappropriate only to rewrite drafts without receiving any feedback on their revision work. Therefore, it was decided to provide students with teacher feedback regarding their achievement; this feedback is not for correcting the second draft but rather serves as a midpoint summative evaluation before moving on to the next session. Figure 2.5 presents the blended design adopted for this study based on the discussion of previous studies.

Figure 2.5

The New Blended Design in the Present Study



Student draft 1 revision ===> Student draft 2 (same text) ===> Student writes text (new text)

Teacher corrective feedback

Teacher feedback

Third, the 19 items on diagnostic assessment proposed by Alderson (2005), which was discussed in Section 2.3, are suggestive in many respects and can be fully utilized in formulating the hypotheses for the present study. In particular, the focus on weaknesses rather than strengths, detailed feedback, integration of SLA and linguistics research, assessment without bias toward vocabulary or grammar, recommendation of computer assistance, etc., form the basic concepts of this study. However, there is one point of Alderson that the present study does not follow: the possibility of using indirect writing tests such as DIALANG to focus on evaluation consistency and the effectiveness of immediate feedback. As Knoch (2007b) suggests, “an indirect test for diagnostic tests lacks face validity and has fallen out of favor in general” (p. 14). Thus, the indirect approach is not adopted in the present study. To sum up, diagnostic feedback based on the formative diagnostic assessment incorporating the concepts proposed by Alderson except for the idea of indirect test is adopted in the present study.

Section 2.4 provided a review of the conceptual definition of coherence in writing. The most significant among them is the idea of rhetorical and ideational focus of the judgment strategies in the taxonomy presented by Cumming et al. (2001, 2002) by defining L2 writing components through raters' decision-making behavior. Moreover, the concept of coherence breaks/anomalies (e.g., Corder, 1967; Knoch, 2007b; Maimon & Nodine, 1978; Wikborg, 1990) in L2 writing assessment is a key issue in EFL writing assessment. One way to overcome the above-mentioned difficulties in organization assessment is to examine the learners' writing closely to determine where the problems lie and to investigate each characteristic, which could be reflected in teacher feedback.

Another focus of the present study is graphics displays in language learning (reviewed in Section 2.5). According to the visual argument hypothesis, "graphical representations are effective because, owing to their visuospatial properties, their processing requires fewer cognitive transformations than does text processing and does not exceed the limitations of working memory" (Vekiri, 2002, p. 281). Furthermore, Winn et al. (1991) found that solving a problem of the students provided with a tree diagram resulted in less time on task. Therefore, the hypothesis is that presenting graphical feedback requires less time on task comparing with presentation of text only. This is directly related to the effectiveness of schematized feedback.

Lastly, based on reviews (Weigle, 2015; Knoch, 2007b) on the rating scales for writing that fit the research objectives, several points should be valued when developing or choosing a rating scale. According to Weigle (2002, pp. 122–124), the following issues should be considered: (a) Who is going to use the rating scale? (b) What aspects are most important and how will they be divided up? In other words, what are the criteria based on? (c) How many points, or scoring levels, will be used? (d) How will scores be reported? The hypothesis is that an analytical rating scale that allows detailed evaluation of student writing from different evaluation perspectives is beneficial. To be more specific, a desirable rating scale should contain descriptors that are easily understood by students, and at the same time, it should be constructed theoretically based on an academic perspective. In other words, an analytical evaluation instrument that is theory-based and student-oriented is desirable.

The above matters are summarized in the RQs 1–3. Regarding the AUA framework (Bachman & Palmer, 2010), RQ 1 addresses Claim 4, RQ 2 addresses Claim 3, and RQ 3 addresses Claim 1. RQ 1, RQ 2, and RQ 3 (1) and (2) are based on quantitative analyses, while RQ 3 (3) and (4) are based on qualitative analyses.

RQ 1: Do the rating scale and the text annotation used in the present study function properly?

RQ 2: How can students' overall writing performance and organization of their writing be characterized/interpreted through the rating scale and the annotation scheme?

RQ 3: To what extent and how does the type of teacher feedback (conventional versus graphic) affect (1) overall writing performance and organization of their writing across occasions (initial draft, revised draft of the initial task, and a transfer task) in terms of scores, (2) revision time on task, (3) ideational and rhetorical coherence in the transfer task (through the analysis of information obtained by the annotation scheme), and (4) students' rewriting behaviors and perceptions?

Chapter 3

Methodology

3.1 Student Participants

A total of 50 second-year university students voluntarily participated in the study. They were enrolled in the College of Education at a private university in Tokyo, Japan. They were from two EFL academic writing classes. Both classes were taught by the author. Of the 50 participants, five students were excluded from subsequent analyses because they missed class and were unable to complete at least one writing task in the classroom and turned it in as take-home assignment. Data from the remaining 45 students were analyzed for this study. Of the two classes, one class (22 students) participated as the control group and the other class, 23 students, participated as the intervention group. In the control group, 13 students (59%) were female, and nine students (41%) were male, while in the intervention group, 13 students (57%) were female, and 10 students (43%) were male. The choice of the control and intervention groups was random. In the morning class that the author was in charge of, the first period was the control group, and the second period was the intervention group. Neither group included English-speaking returnees. Moreover, none of the students were from domestic international schools.

Figures 3.1 and 3.2 show a comparison between the control and intervention groups regarding the scores on the Eiken test (EIKEN Foundation of Japan) and the TOEIC Listening and Reading (L&R) Test at the beginning of the course. Most of the target students had taken these two standardized English language proficiency tests, and only a very limited number of students had experience taking other tests such as the TOEFL iBT. There was a difference between the groups, with the intervention group having a higher percentage of Grade 2 holders than the control group. No students were Grade Pre-1 in the intervention group. However, looking at the TOEIC L&R results, there was not a difference between the groups.

Figure 3.1

Eiken Grades in the Control and Intervention Groups

Control Group (*n* = 22) Intervention Group (*n* = 23)

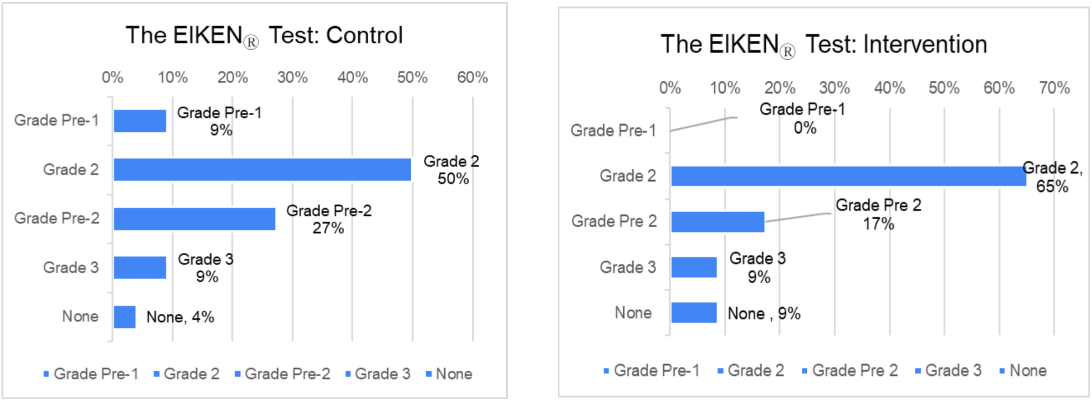
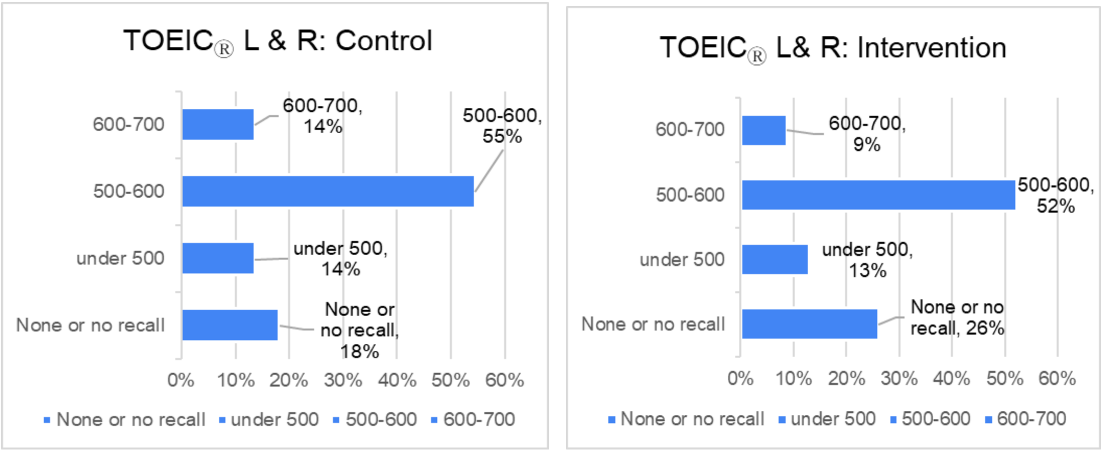


Figure 3.2

TOEIC L&R Score Distribution for the Control and Intervention Groups

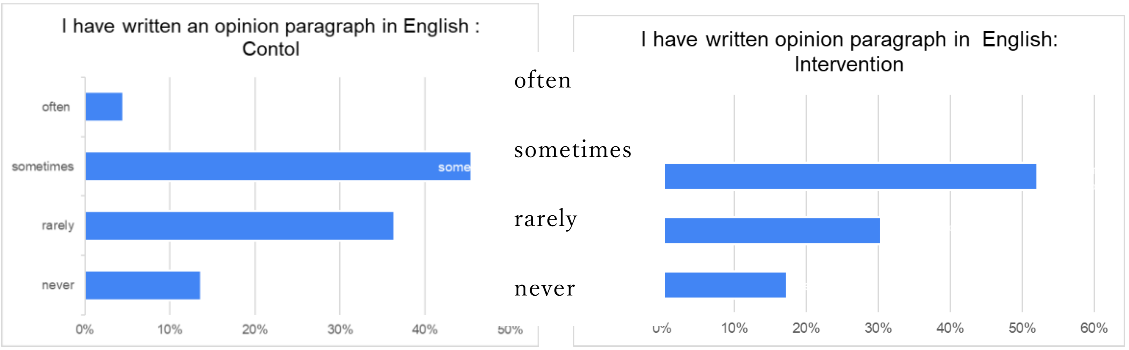
Control group (*n* = 22) Intervention group (*n* = 23)



Figures 3.3 and 3.4 show the students' previous experience related to opinion paragraph writing in English. Both the control and intervention groups had had experience writing opinion paragraphs in English but had rarely or never received instruction on how to write them in class. Opportunities to write opinion paragraphs in English seemed to be related to preparation for the Eiken writing test, which started at all levels in 2017.

Figure 3.3

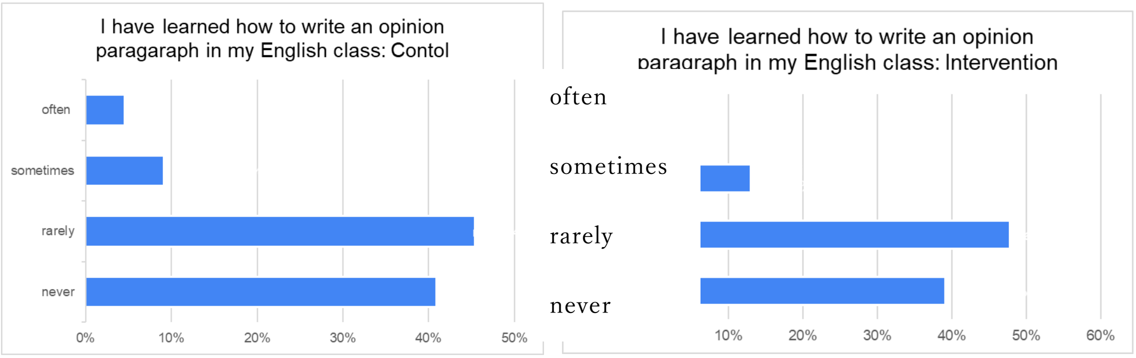
Opportunities to Write Opinion Paragraphs in English



Control group (n = 22) Intervention group (n = 23)

Figure 3.4

Opportunities to Learn How to Write Opinion Paragraphs in English in the Classroom



Control group (n = 22) Intervention group (n = 23)

3.2 Materials

3.2.1. Writing Tasks

Table 3.1 presents a list of writing prompts employed in the present study. They were adopted from the Eiken Grade 2 and Pre-1 tests. These tests were chosen as the topics for this study because a preliminary survey revealed that most of the target students had English proficiency at the Eiken Grade 2 level and were aiming for the Grade Pre-1 level. The reason for choosing a topic from the Eiken writing test is that the test, like this study, requires a statement of opinion, which can be written in about one paragraph. The Grade 2 topic is a social one from a personal perspective, while the

Grade Pre-1 test requires writing coherent text on a highly social topic from a more objective perspective, and thus it is considered appropriate for intermediate-level university students. Furthermore, from a research perspective, two tasks were selected from different grades to set up different levels of difficulty. The last task to be worked on was the transfer task, which was assigned a Grade Pre-1 level question. This was selected out of consideration that as a finishing exercise, the students should complete a topic at the level of difficulty for which they were aiming.

Table 3.1

List of Writing Prompts

Time	Prompts for both control and intervention groups (counter-balanced)	Source
Pretest (initial)	(A) There is a view that young people should spend more time thinking about their future careers. Do you agree with this opinion? Give your opinion about this topic. （若者は自分の将来のキャリアについて考える時間を増やすべきだという意見があります。この意見に賛成ですか。あなたの意見を述べなさい。） (B) There is a view that big companies have a positive effect on society. Do you agree or disagree with this view? Give your opinion about this topic. （大企業は社会に良い影響を与えているという意見があります。この意見に賛成ですか、反対ですか。あなたの意見を述べなさい。）	(A) Eiken Grade 2 (2020-2) (B) Eiken Grade Pre-1 (2021-1)
Posttest (revision)		
Transfer task (retention)	(C) Is it beneficial for workers to change jobs often? Write your opinion to answer this question. Although this is a Y/N question, don't start your passage with Yes or No. (労働者が転職を繰り返すのは得/有益ですか。この質問に答えてあなたの意見を書きなさい。ただし、Yes/No 疑問文ですが、文は Yes/No で始めないこと。）	(C) Eiken Grade Pre-1 (2021-2)

In selecting topics for the study, consideration was given to the content compatibility of "general social issues related to working in society after college," which is a common interest of these college students. When selecting topics, Grade 2 was chosen from the 2020 test and Grade Pre-1 was selected from the 2021 test, in order to avoid questions on the tests that the students had been taken previously. As a sample of instruction used for the pre- and post-task Topic A are presented in Appendices A and B, respectively. The instructions are written in Japanese, and the topics are accompanied by a Japanese translation by the author to avoid students' misunderstandings of the prompt.

For the pre-tasks, the following seven points were presented in Japanese on the sheet as the initial task directions: (a) State you opinion with two (or more) reasons to support it. (b) Write "TOPIC A" in the title. (c) The word count should be at least 100 words, and about 120 words is preferable. You may exceed 120 words. (d) Points may be deducted if the word count is not sufficient. (e) You are encouraged to write by incorporating "opposing viewpoints" (counterarguments, rebuttals) taught in class. (f) The time limit is 30 minutes (strictly adhered to). (g) You are not allowed to use dictionaries or Internet searches (see Appendix A).

For the post-tasks, which is revision writing, no time limit was set, and the students were instructed to submit it to the web-based platform as soon as it was finished. In the revision process, no dictionaries or cell phones were allowed. The students were instructed to review and rewrite their initial draft with the teacher feedback that had been returned to them just prior to the occasion. The following are directions for the post tasks: (a) Refer to your teacher's feedback and revise your paragraph. (b) Feel free to revise not only what your teacher pointed out, but also what you think you should revise, (c) Write "TOPIC A (Revised)" in the title. (d) the word count should be at least 100 words, and about 120 words is preferable. You may exceed 120 words, (e) There is no time limit. Submit your revisions to X (the name of the university's web platform) when you are finished, (f) Do not use a dictionary or Internet search, (g) For your reference, the topic is reposted below (see Appendix B).

In the actual Eiken test, three key words are provided as possible reasons for each prompt, but not in this test. This is because the work time is longer than that of the Eiken test. Moreover, there is more time to spend on planning than for the Eiken test.

3.2.2 Rating Scale

3.2.2.1 The Base Rating Scale. For scoring primary data (students' paragraphs written under the pre-task, post-task, and retention conditions), the ESL CP (Jacobs et al., 1981) was adapted for the present study. This scale was chosen because it is suited to profile an individual's writing skills with five clearly defined sub-components with differential weighting: content, organization, vocabulary, language use, and mechanics. The highest possible total score is 100, but it is possible to analyze each sub-category independently. Above all, the analytic rating scales with detailed descriptors are consistent with this study's focus on diagnostic assessment. The scale is an analytical assessment rubric for ESL writing that has been widely used for many years and is particularly well-established for use in classroom writing assessments (e.g., Barkaoui, 2007; Cumming, 2009; Weigle, 2002). Barkaoui (2007), in particular, considers the scale as one of those which are "appropriate for marking argumentative essays by EFL university students" (p. 90). The scoring sheet attached in their book was used as one of the writing feedback materials in the present study.

3.2.2.2 Rating Scale in the Present Study. This study employed a modified version of the ESL CP with revisions to the study specifications. There are three main revisions: changes to the weights assigned to the rating criteria, revision of level scores, and some modifications of the descriptors.

3.2.2.2.1 Modification of the Analytic Rating Scale. Because the distribution of scores for the ESL CP is weighted toward content and language, the distribution of the scores was adjusted considering the purpose of this study. The weight of organization was increased from 20% to 25% and the weight of content was reduced by 5%, to 25%, so that the two were in the same ratio. This change more appropriately reflected the objective of this study, namely the organizational aspect of writing as well as the content.

The other parts of the weighting remain unchanged. Table 3.2 shows the modified score distribution by criteria for the present study based on ESL CP. The next sections describe the modifications and the rationale for them.

Table 3.2

Modification of the Score Allocation for Each Criterion

Focus	Criterion	ESL CP Original	Modified	Subtotal
Rhetorical and Ideational focus	*Content*	30	25	50
	Organization	20	25	
Language focus	*Language use*	25	25	50
	Vocabulary	20	20	
	Mechanics	5	5	
Total		100	100	100

Note. Score changes are underlined.

As a result, the combination of content and organization corresponding to the "rhetorical and ideational" aspects, and the combination of the remaining three corresponding to the "language focus" aspects (Cumming et al., 2000, 2001) have an equal weight, which has not changed the original ESL CP allocation.

3.2.2.2.2 Conversion of Raw Scores Into Level Scores. Table 3.3 shows the correspondence between the newly developed 6-point rating scale and the four-level ordinal scale of the ESL CP accompanied by the corresponding score range for each level. The raw scores produced in each category in the ESL CP were converted to the 6-point scale for further analyses. The exception was the mechanics score, which initially had a small score range of 5 points. Accordingly, the original scores were treated as level points, and no score conversion was necessary. Vocabulary is the only criterion that has a maximum score of 20 points, so the distribution of points by level is different from the other three criteria (see Table 3.3)

Table 3.3

Newly Developed 6-Point Scale Compared With the Original Four-Level Ordinal Scale of the ESL CP

ESL CP original scale			Newly developed scale		
Score range		Ordinal four levels	Six levels	Subdivided score range	
22–25	(18–20)	Excellent to very good	6	above 22	(above 18)
18–21	(14–17)	Good to average	5	20–21	(16–17)
			4	18–19	(14–15)
11–17	(10–13)	Fair to poor	3	15–17	(12–13)
			2	11–14	(10–11)
5–10	(7–9)	Very poor	1	under 10	(under 9)

Note. The numbers in parentheses are the scores for the *Vocabulary criterion with a full score of 20.*

As shown in Table 3.2, in the original ESL CP a 100-point scale and four levels of evaluation criteria coexist. The four-level ordinal scale includes Excellent to very good (highest level), Good to average, Fair to poor, and Very poor (lowest level). It makes sense to provide a more granular measure of the level implied by the score than simply the score itself for both students and raters. However, it was determined that it would be reasonable to convert the four levels into a 6-point rating scale. The following two points are the rationale for this decision.

First, the author's follow-up discussions with the raters revealed that in the rater's scoring decision-making process, he/she first determined the level on the four-ordinal scale before determining a specific score. He/she then decided whether it belongs to the upper or lower half of that score range to determine their final score. The raters also remarked that this process was required particularly for the middle two levels, while determining the highest and lowest levels was relatively easy. This implies that the raters virtually rated the students' writing samples at six levels for scoring. The author felt exactly the same way when she participated as one of the raters. As Barkaoui (2007) reported in his study on the rating scale impact on EFL essay marking by raters, they actually used a "self-generated" or self-interpreted way of using the rating scale to some extent even in the analytic rating scale accompanied by descriptors. It is important for the validity and reliability of the rating scale to be as close as possible to the reality of the rated behavior. While the fine-grained 100-point scale has the sensitivity to capture small growths and differences among learners, one is faced with the dilemma of whether a 1-point difference within the same range is meaningful.

The other rationale for developing a 6-point scale is that the majority belonged to the two intermediate levels, which raises the concern that the granularity at the four levels is not adequate.

3.2.2.2.3 Modification of the Descriptors in the Rating Scale. The descriptors in the ESL CP were modified for the present study to score one-paragraph writing samples because the ESL CP was originally designed to evaluate essay writing. Some phrases were added for clarification of the wording to help raters to have a better common understanding. Moreover, in terms of content, some descriptors were added to reflect evaluation of the argument text, which is the focus of this study. To be specific, the organization criterion was modified to include a description of elements such as counterargument and rebuttal featured in this study. The revised parts of the organization level descriptors are presented in Table 3.3. A Japanese translation is also attached along with the English (Appendix C). The Japanese translation is intended to provide a common understanding of the concepts involved in the evaluation. This is because the primary language of the raters in this study is Japanese. Moreover, the definition of terms is not always straightforward. This modification is a further revision to the original form described by Matsumura and Sakamoto (2021).

Table 3.4

Descriptors and Criteria for the Organization Scale for Paragraph Writing Focusing on Argumentation Based on ESL Composition Profile (Jacobs et al., 1981, p.93; Matsumura & Sakamoto, 2021, p.50)

Descriptor	ESL CP criteria in question	Modified and/or added parts in the study
Fluent expression	Are there introductory and concluding paragraphs? Are there effective transition elements - words, phrases, or sentences - which link and move ideas both within and between paragraphs?	Are there introductory and concluding **sentences?** Are there effective transition elements - words, phrases, or sentences - which link and move ideas **within the paragraph**? ***Are there any abrupt or unintelligible sentences that interrupt the flow?***
Ideas clearly stated/supported	Is there a clearly stated controlling idea or central focus to the paper (a thesis)? Do topic sentences in each paragraph support, limit, and direct the thesis?	Is there a clearly stated controlling idea or central focus to **the paragraph** (**a topic sentence**)? Do **sentences** support, limit, and direct the ***topic*** **sentence**?
Succinct	Are all ideas directed concisely to the central focus of the paper, without digressions?	Are all ideas directed concisely to the central focus of **the paragraph**, without digressions? ***Does it contain repetitive or redundant sentences?***

Continued

Descriptor	ESL CP criteria in question	Modified and/or added parts in the study
Well-organized	Is the overall relationship of ideas within and between paragraphs clearly indicated? Is there a beginning, a middle, and an end to the paper?	Is the overall relationship of ideas within and between ***sentences*** clearly indicated? Is there a beginning ***(a topic sentence)***, a middle ***(a group of supporting sentences)***, and an end ***(conclusion or restatement)*** to the writing? ***Are elements of awareness of different viewpoints or the limitations of his or her argument, such as counter argument and rebuttal, incorporated into the support section?***
Logical sequencing	Not applicable	***Does the element of awareness of different viewpoints develop in a rational manner?***
Cohesive	Does each paragraph reflect a single purpose? Do the paragraphs form a unified paper?	Does each **group of sentences** ***(the topic sentence, major points, and the concluding sentence)*** reflect a single purpose? Do the **group of sentences** form a unified writing?

Note. The phrases in bold are the parts that were modified and those in italics are the newly added part.

3.2.3 Raters and Task Assignment

3.2.3.1 Raters. The students' writing samples were evaluated by six raters (R1–R6) in their 30s–50s who teach EFL classes, including English writing, at universities in Japan. Their teaching experience varied from 3 to 24 years. L1 Japanese speakers were chosen as raters because the university where the target students are enrolled provides instruction in beginning to intermediate academic writing by an L1 Japanese instructor. Therefore, having L1 Japanese raters is consistent with the context of the classroom. However, five of them have had schooling in English-speaking countries, and four have had previous university education in the United States or the United Kingdom. Table 3.4 summarizes the raters' academic background at the time of their study participation.

Table 3.5

Information on the Raters

Rater	Teaching experience in university	Highest academic degree earned	Major research interest
R1 (Female)	11 years	MA	Applied linguistics, intercultural communication
R2 (Male)	3 years	MA	English education, language testing
R3 (Female)	5 years	PhD	English education, phonetics
R4 (Female)	9 years	MA	L2 socialization, English-medium instruction
R5 (Female)	24 years	MA	EFL writing
R6 (Female)	5 years	MA	English education, EFL writing, language testing

Note. The above information refers to the time the study was conducted (in 2021).

3.2.3.2 Rater Training Sessions and Post-Evaluation Interviews. Prior to the students' writing evaluation, the raters completed three rater training sessions totaling 6 hours in duration. The session was an interactive online training session, which brought everyone together. The session consisted of three sections: an overview of the research objectives and student background, an introduction to the assessment instrument and specific operationalization, and a presentation of benchmark student writing and an actual rating exercise. During the actual rating exercise, the results of the evaluation were presented to the participants, and the scoring criteria were confirmed and reconciled based on a discussion. In addition, the raters watched a 1-hour prerecorded on-demand video of the author's explanation for a review of the rating scale and reminders accompanied by more exercise samples with benchmarks. As a follow-up post-evaluation interview, each rater had an individual online discussion with the author that lasted for approximately 1 hour. The discussion focused on the individual students' writing, which the author determined required confirmation based on the rating results. Furthermore, they were interviewed about how they actually implemented the rating scale or about the evaluation process in general.

3.2.3.3. Rater Task Assignment. Table 3.6 shows the rater assignment for writing tasks. Each of the six raters was assigned to evaluate the same student's pre-task and post-task writing samples on one of the three topics (Topics A, B, and C). Scoring student writing samples on each topic involved three raters. R2 and R6 was involved in all evaluations as the second rater. For Topic C (the transfer task), the author and another rater (R5) evaluated all 50 writing samples. R5 is also one of the annotators who used the web-based TIARA annotation tool (discussed in Section 3.2.5). This combination of assignments resulted in two ratings per student response.

Table 3.6

Rater Assignment for Writing Tasks for Evaluation

Rater	Topic A				Topic B				Topic C			
	Pre-A and post-A				Pre-B and post-B				Transfer task			
	Control		Intervention		Control		Intervention		Control		Intervention	
	1st half	2nd half	1st half	2nd half	1st half	2nd half	1st half	2nd half	1st half	2nd half	1st half	2nd half
R1	x		x						(x)		(x)	
R2		x		x					(x)		(x)	
R3					x		x		(x)		(x)	
R4						x		x	(x)		(x)	
R5									x	x	x	x
R6	x	x	x	x	x	x	x	x	x	x	x	x

Note. The "x" represents the section for which the rater was responsible. A "(x)" indicates that the rater was responsible for some of the writing for inter-rater agreement check.

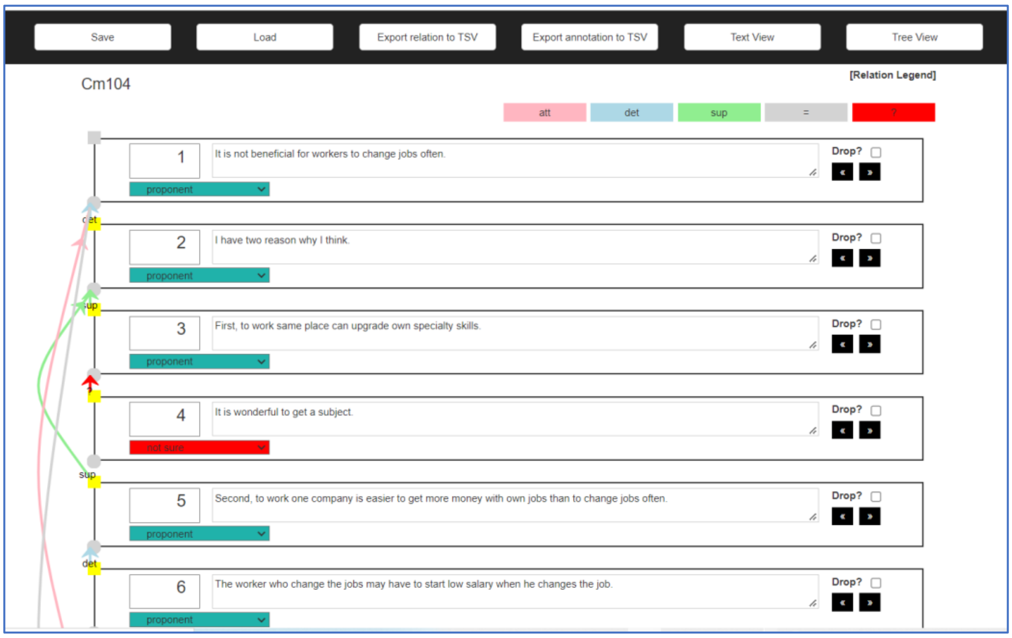
Additionally, to examine the degree of agreement in scoring among the six raters, an exact match of five student written responses from the first half of Topic C was checked, where exact agreement was 79%, adjacent agreement was 18%, and non-adjacent agreement was 3% for the six raters. The final scores given to each student response were the mean across the two ratings for each scoring rubric.

3.2.4 Annotation Tool

As described in Chapter 1, students' paragraph writing samples were analyzed for feedback generation with a customizable web-based annotation tool called TIARA (Putra et al., 2020). The tool was developed by a research group from the Tokyo Institute of Technology. It is considered to be useful for educational purposes as well both in learning-to-read and learning-to-write scenarios (TIARA Manual, 2021). The novelty of TIARA lies in the dual-view user interface (the text and tree views) shown in Figures 3.5 and 3.6, which provides the annotators with a global overview of the text to examine their annotation results and make local changes when necessary. Figure 3.5 is a screenshot of the TIARA interface showing a working area for tagging and linking the sentences, where numbered segmented text units are presented.

Figure 3.5

Screenshot: Tagging and Linking Sentences in TIARA



Cm104

1 It is not beneficial for workers to change jobs often. proponent Drop? □

2 I have two reason why I think. proponent Drop? □

3 First, to work same place can upgrade own specialty skills. proponent Drop? □

4 It is wonderful to get a subject. not sure Drop? □

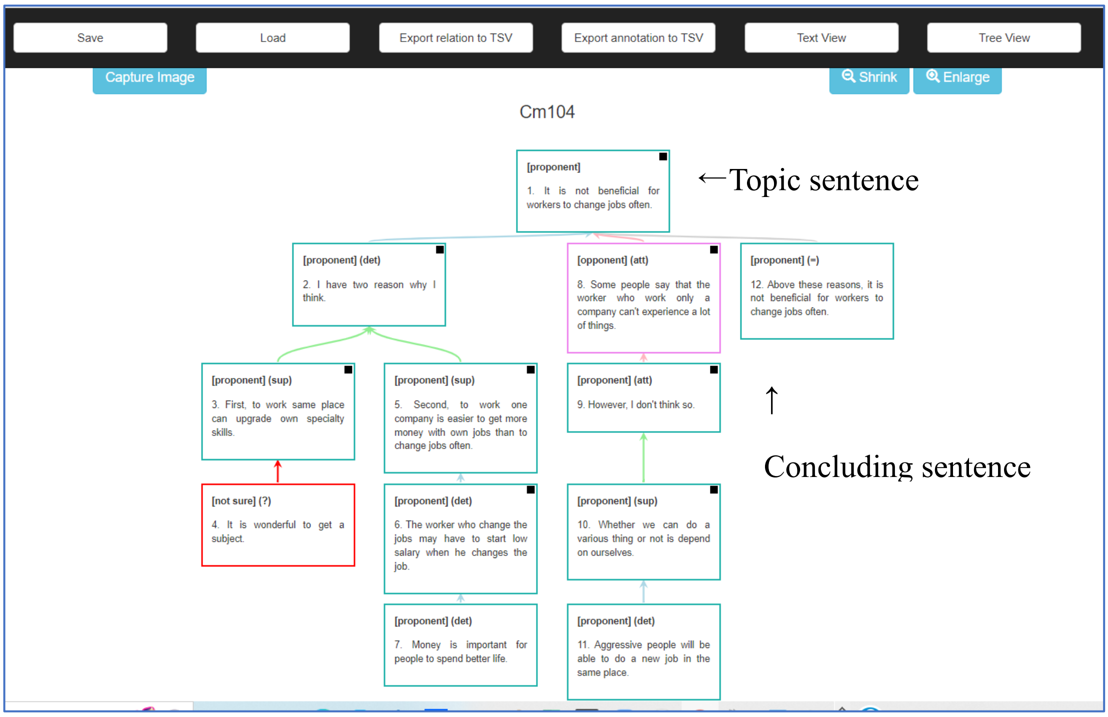
5 Second, to work one company is easier to get more money with own jobs than to change jobs often. proponent Drop? □

6 The worker who change the jobs may have to start low salary when he changes the job. proponent Drop? □

Figure 3.6 shows a tree-shaped diagram automatically converted from the composition of the entire text that was annotated by an annotator. It visualizes logical sequencing between text units in a simple layout. The original text unit is presented in each box so that users can see the actual sentence and check the overall structure at a glance. The "capture image" function allows to create a screenshot, which allows users to share the view with others.

Figure 3.6

Screenshot: A Tree-Shaped Visualization of the Annotated Diagram Generated by TIARA



Cm104

[proponent]
1. It is not beneficial for workers to change jobs often.

←Topic sentence

[proponent] (det)
2. I have two reason why I think.

[opponent] (att)
8. Some people say that the worker who work only a company can't experience a lot of things.

[proponent] (=)
12. Above these reasons, it is not beneficial for workers to change jobs often.

[proponent] (sup)
3. First, to work same place can upgrade own specialty skills.

[proponent] (sup)
5. Second, to work one company is easier to get more money with own jobs than to change jobs often.

[proponent] (att)
9. However, I don't think so.

↑

Concluding sentence

[not sure] (?)
4. It is wonderful to get a subject.

[proponent] (det)
6. The worker who change the jobs may have to start low salary when he changes the job.

[proponent] (sup)
10. Whether we can do a various thing or not is depend on ourselves.

[proponent] (det)
7. Money is important for people to spend better life.

[proponent] (det)
11. Aggressive people will be able to do a new job in the same place.

Although TIARA has more functionality to edit, reorder, and delete/drop text, the present study did not take advantage of these features, because the learners were asked to perform these modifications themselves. This study used the following three functions: classifying sentences into their rhetorical categories, *proponent*, *opponent*, or *not sure*, by selecting from the drop-down menu; establishing a connection between two units by dragging an arrow from the source to the target unit; and establishing a relation by choosing the label, that is, relation labels, the user has created beforehand (see Table 3.6). The labels in Table 3.6, except for "?" are set as default in TIARA based on the components of the Toulmin model of argument (Toulmin, 2003); the user can add or change labels with a simple modification of the program. In this study, the relation label "?" (questionable) was added, which indicates a coherence anomaly. The description in the table is adapted from Matsumura and Sakamoto (2021, p. 36).

Table 3.7

Relation Labels Used in this Study

Label	Name	Description
sup	Support	The source sentence asserts or justifies reasons and ideas for supporting the target sentence. It roughly corresponds to "evidence" in RST.
det	Detail	The source sentence further explains, describes, elaborates or provides background for the concept(s) mentioned in the target sentence. It roughly corresponds to a combination of "elaboration" and "background" in RST.
Att (opponent)	Counter-argument	The source sentence considers a counterargument that argues for the opposite opinion.
Att (proponent)	Rebuttal	The source sentence considers rebuttals for the counterargument, either arguing against it or indicating circumstances when the main argument does not hold true.
=	Restatement	The sentence summarizes important parts of the main argument for the second time. Restatements are directly connected to the topic sentence.
?	Questionable	The annotator has difficulty in identifying the relation between the source sentence and any previous sentence. It suggests some coherence breaks.

Note. The descriptions are adapted from Matsumura and Sakamoto (2021, p. 36).

Table 3.7 summarizes the link labels used in the study: support (sup), detail (det), attack (att), restatement/conclusion (=), and questionable (?). The label questionable was added by the authors for the sentences for which the annotators could not identify the label to address the link, which are categorized as “anomalies” for further analysis. This information on relation labels will be used for qualitative discourse analysis as well as investigation of inter-annotator agreement along with that of the combination of pairs of units, which is explained in the next paragraph.

The results of annotation are saved and exported to the spreadsheet-friendly TSV format, which presents extracted information on relations of all possible pairs of units. This is useful for calculating inter-annotator agreement by checking the agreement of the combination of source unit and target unit.

3.2.5 Annotators

Two annotators annotated the students’ writing samples with TIARA; they also served as raters R5 and R6 (see Table 3.5). The two had collaborated for several years, including pilot studies using the annotation tool employed in this study. Putra et al. (2021), the creators of TIARA, describe the inter-annotator agreement as follows: “Our argument annotation scheme is demonstrably stable, achieving good inter-annotator agreement and near-perfect intra-annotator agreement” (Putra et al., 2021, p. 1).

The inter-annotator consistency was measured by two indices: source-target agreement and relation label agreement. The agreement statistics were calculated with the data of 45 writing samples on the transfer task (Topic C). From these 45 writing samples, a total of 453 text units were obtained. Of these, source-target discrepancies were identified in 63 text units, and relation-label discrepancies were identified in 88 text units. The agreement for the former, that is, to which target unit the source unit leads, was 0.86, and the agreement for the latter, that is, to which relation the source unit was labeled, was 0.81. The two agreement scores calculated previously for another study sample with the same two annotators were 0.81 and 0.75, respectively. Again, the agreement on the source–target text unit links tended to be higher. The agreement in the annotation of 30 writing samples randomly selected out of 50 stand-alone paragraphs written by high school students was checked (Matsumura & Sakamoto, 2021).

3.2.6 Teacher Feedback (Control and Intervention Groups)

In this study, two different types of teacher feedback were prepared based on student writing samples: conventional corrective feedback for the control group, and schematized feedback with a tree-shaped diagram created with the TIARA annotation tool for the intervention group. In both classes, the author was the instructor, and all teacher feedback was generated by the author. The form of feedback is shown in Table 3.8.

Table 3.8

Types of Feedback Given to the Students

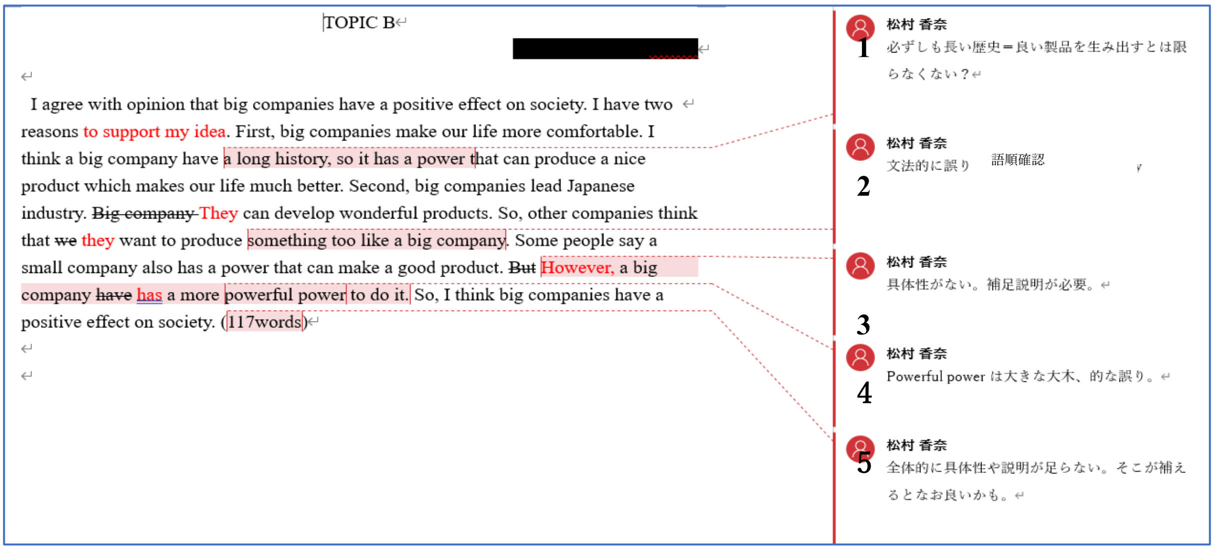
	Types of feedback	
	Delayed teacher written feedback (Input-providing & output-pushing)	
Group	Paper-based	Web-based
Control	**Conventional text-based feedback** (a mixture of direct and indirect plus metalinguistic feedback)	・Overall comment (metalinguistic feedback) ・Scores on the analytic rating scale
Intervention	**Schematized feedback** (with a mixture of direct and indirect feedback)	

The feedback content was the same for both groups except where bolded in Table 3.7. Delayed written feedback was chosen because it is common in the case of writing, and immediate oral feedback can strain learners' short-term memory and make it difficult for teachers to provide feedback on discourse, such as content and organization (Tanaka, 2015, p. 109). Moreover, while oral corrective feedback has the advantage of being interactive and allowing for individual confirmation of the author's intentions, it also has drawbacks: It is time-consuming for individual interviews and there is a possibility of overlooking or missing something on both the teacher's and the student's sides. In addition, from a research perspective, delayed written feedback was chosen to control the content and quantity of feedback. In the written corrective feedback, direct feedback often explicitly indicates the student's written errors along with suggested revisions (Ellis, 2010). This type of feedback often refers to language use, such as mechanics, spelling errors, and obvious misuse of vocabulary, which strongly suggests output-pushing, that is, revising the text. (Ashwell, 2000; Biber et al., 2011). On the other hand, indirect feedback is pointing out errors by underlining or coding, inexplicitly indicating the problems. Along with this type of feedback, feedback on content or organization is sometimes regarded as input-providing, which means whether or not the learner decides to accept the input may be optional. However, it should be noted that any form of written feedback on writing is explicit in terms of the written indication (e.g., Ashwell, 2000; Bitcher & Storch, 2016; Storch, 2010; Ellis, 2010; Tanaka, 2015; Zamel, 1985).

Figures 3.7 and 3.8 show examples of actual teacher feedback for both groups. The instructor (the author) tried, where possible, to avoid group bias, although the number and content of feedback comments required would vary with each student's writing sample. Furthermore, different forms of feedback as these may require different comments. However, the instructor (the author) tried as much as possible to avoid group bias in the comments. The comments were written in Japanese, and an English translation is provided for this paper. The English comments are not shown to students. The type of comment is noted in square brackets after the translated comment.

Figure 3.7

A Sample of Text-based Feedback Sheet (Control Group)



TOPIC B

[redacted]

I agree with opinion that big companies have a positive effect on society. I have two reasons to support my idea. First, big companies make our life more comfortable. I think a big company have a long history, so it has a power that can produce a nice product which makes our life much better. Second, big companies lead Japanese industry. ~~Big company~~ They can develop wonderful products. So, other companies think that ~~we~~ they want to produce something too like a big company. Some people say a small company also has a power that can make a good product. ~~But~~ However, a big company ~~have~~ has a more powerful power to do it. So, I think big companies have a positive effect on society. (117words)

1 松村 香奈
必ずしも長い歴史＝良い製品を生み出すとは限らなくない？

2 松村 香奈
文法的に誤り　語順確認

3 松村 香奈
具体性がない。補足説明が必要。

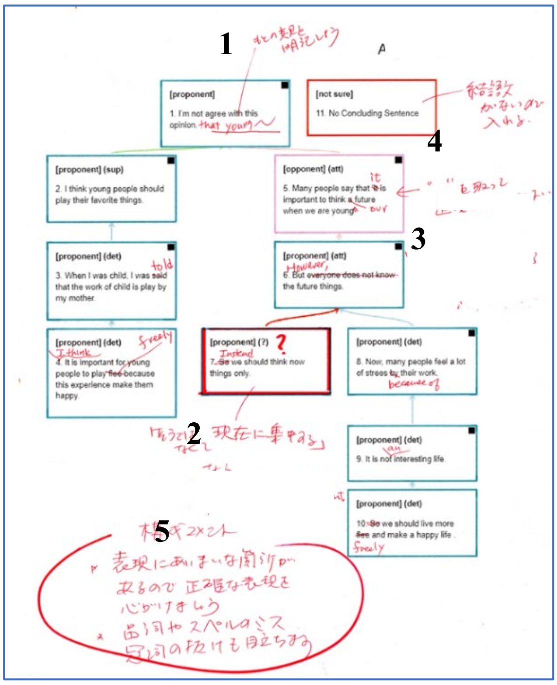
4 松村 香奈
Powerful power は大きな大木、的な誤り。

5 松村 香奈
全体的に具体性や説明が足らない。そこが補えるとなお良いかも。

[Translation of the comments above]
1. Doesn't a long history necessarily produce a better product? [content]
2. Grammatically wrong. Check the word order.[language]
3. No specifics. Elaborate more. [content]
4. "Powerful power" sounds like huge big tree.[vocabulary/language]
5. Overall, there is a lack of specificity and explanation. It would be better if there were.
(Several more corrections and additions are in the text.)

Figure 3.8

A Sample of Schematized Feedback Sheet (Intervention Group)


1
主張を明記しよう
A
[proponent]
1. I'm not agree with this opinion. that young
[not sure]
11. No Concluding Sentence
結論文がないので入れる。
4
[proponent] (sup)
2. I think young people should play their favorite things.
[opponent] (att)
5. Many people say that it is important to think a future when we are young our
" " を取って
3
[proponent] (det)
3. When I was child, I was said told that the work of child is play by my mother.
[proponent] (att)
However, 6. But everyone does not know the future things.
[proponent] (det)
I think 4. It is important for young people to play free freely because this experience make them happy.
[proponent] (?)
Instead 7. So we should think now things only.
[proponent] (det)
8. Now, many people feel a lot of strees by because of their work.
2
「代わりに 現在に集中する」
[proponent] (det)
9. It is not an interesting life.
[proponent] (det)
10. So we should live more free freely and make a happy life.
5
表現にあいまいな箇所があるので正確な表現を心がけましょう
品詞やスペルのミス 冠詞の抜けも目立ちます

[Translation of the comments above]
1. Clearly state the statement of claim.[content/organization]
2. Does it mean "Instead, we should focus on" ? [content]
3. How about deleting double quotation and use direct speech? [mechanics]
4. No concluding sentence. [organization]
5. Please be precise as there are many ambiguous expressions throughout. There are errors in terms of parts of speech, omitted articles, and misspellings. (Several more corrections and additions are in the text.)

3.2.7 Questionnaire on Revision and Teacher Feedback

The students were also administered paper-based descriptive questionnaires (Appendix D). The students were asked two open-ended questions in Japanese and responded in Japanese. The English translation of the questions is: Q.1 What did you fix or modify during the revision process? Q.2 What did you think about the feedback from the teacher? Please write your impressions. In the first question, the participants were asked to describe what they had revised in each of the five analytic criteria separately: content, organization, vocabulary, language use, and mechanics.

The questionnaire is a part of the qualitative data for RQ 3 to investigate the students' behaviors during revising work and the perception on teacher feedback. These are the source data for thematic analysis to explore the effectiveness of schematized teacher feedback.

3.3 Procedure

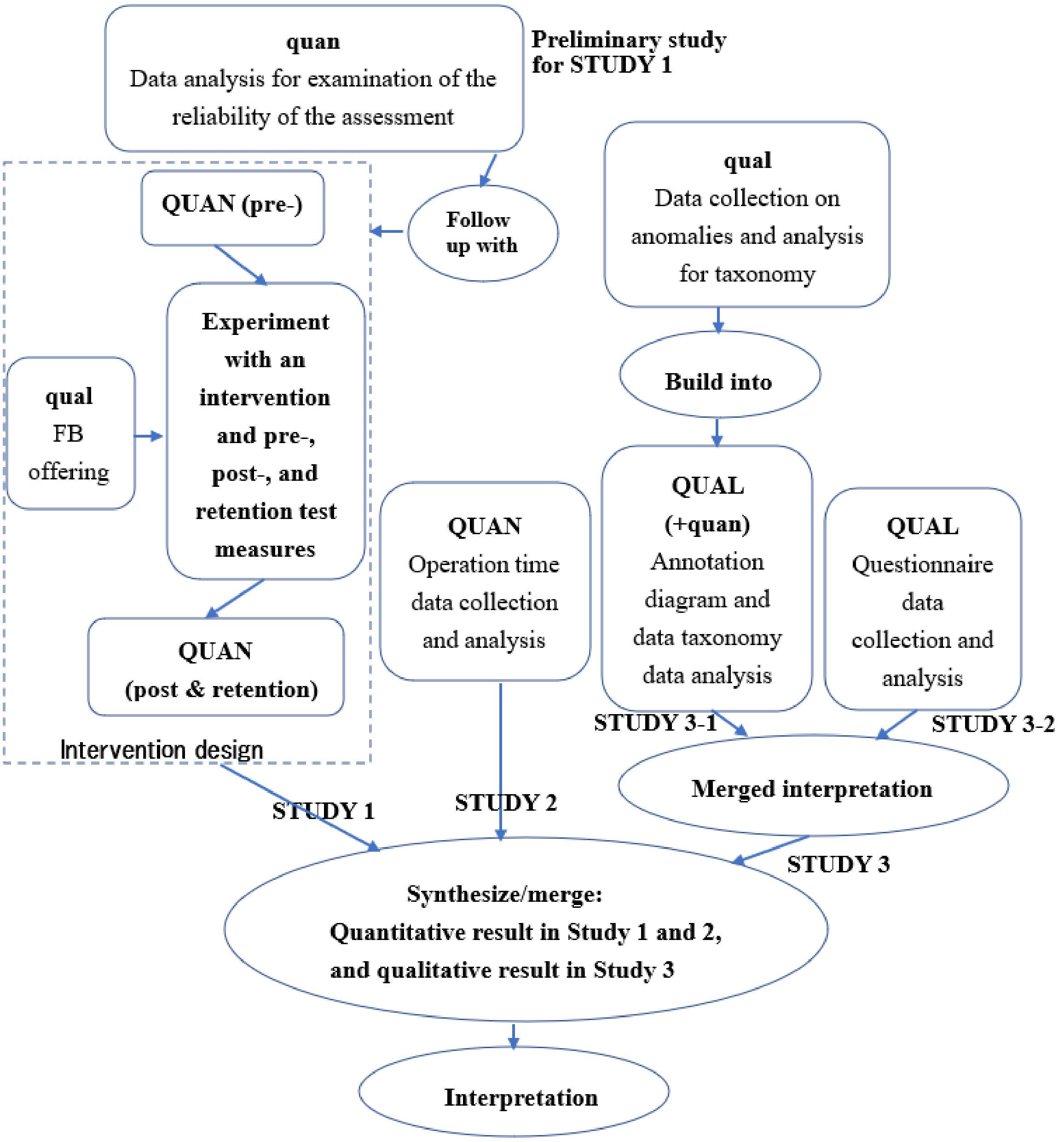
The present study was conducted during the 2021 academic year after undergoing ethical review from the university. Prior to conducting the study, an overview and the purposes of the research was given to the students, explaining that the study would be conducted as part of a class. The students were also informed that they would not be disadvantaged under any circumstances by this study. They were told that the writing pieces and the results of the survey would be subject to analysis for research purposes, and that the respondents' information would be anonymous and coded for possible publication. Although the students were supposed to participate in the activities as part of their regular coursework, their willingness to participate in the research analysis and publication was confirmed when they signed a written consent form.

3.3.1 Research Design

3.3.1.1 Mixed Methods Research Design. Figure 3.9 summarizes the overall mixed methods research design. This investigation consists of three main studies. Study 1 is a quantitative study in terms of scores obtained by the rating scale. In the intervention design, the three writing samples prepared by the students—the initial draft, the revised draft based on the teacher feedback, and the transfer task draft—were compared. The change over three occasions was investigated, and the effectiveness of the schematized feedback was evaluated by comparing the results from students who received schematized feedback with the results from students who received the conventional text-based feedback. The preliminary study of examining the consistency of the rating scale was conducted in advance. Study 2, also a quantitative study, was designed to determine the impact of schematized feedback of time on revision work. Finally, Study 3 consisted of two different qualitative studies. First, Study 3-1 involved analyzing with tree diagrams generated by the annotation tool. This study was intended to address RQ 3. Study 3-2 was conducted by a thematic analysis of the responses from the questionnaire survey of students. The questionnaires submitted to the students addressed RQ 3 (3) and RQ3 (4), their behaviors on revisions and their perception of the teacher feedback. The analytical methods and procedures, and their relationship to the RQs in each study are presented in detail in the sections that follow.

Figure 3.9

The Research Design of the Present Study: A Convergent Mixed Methods Design Using an Intervention


quan
Data analysis for examination of the reliability of the assessment
Preliminary study for STUDY 1
Follow up with
QUAN (pre-)
qual
FB offering
Experiment with an intervention and pre-, post-, and retention test measures
QUAN (post & retention)
Intervention design
qual
Data collection on anomalies and analysis for taxonomy
Build into
QUAN
Operation time data collection and analysis
QUAL (+quan)
Annotation diagram and data taxonomy data analysis
QUAL
Questionnaire data collection and analysis
STUDY 3-1
STUDY 3-2
Merged interpretation
STUDY 1
STUDY 2
STUDY 3
Synthesize/merge:
Quantitative result in Study 1 and 2, and qualitative result in Study 3
Interpretation

3.3.1.2 Establishment of the RQs and Procedure in the AUA Framework. As discussed in Chapter 1, the analysis and discussion of the individual studies in the present research follows the AUA framework (Bachman & Damböck, 2017; Bachman & Palmer, 2010). While a mixed methods approach (Creswell, 2015; Kakai, 2015; Miller & Bustamante, 2016) was adopted as the overall research design, the RQs and procedure were established in the logical AUA framework. In the AUA, the assessment study proceeds sequentially based on the assumption of the previous claim by examining each intended claim. However, the procedure does not necessarily move in one direction; rather, it proceeds cyclically and interactively. Each study, the RQs and claims in the AUA, the mixed methods research perspective, as well as the source data and analysis with methods are summarized in Table 3.9.

3.3.2 The Procedure in Terms of the RQs

After collecting data, various quantitative and qualitative analyses were performed. Table 3.9 summarizes the analysis methods used in this study for each study in terms of the RQs. The left column lists each study, the center column lists the related RQ(s), and the right column lists the analysis method and the data source analyzed. Regarding RQ 2, the aim was to address the meaningfulness in interpreting scores and annotated diagrams, which is the premise of RQ 3 (Claim 1 in the AUA). Therefore, except for the descriptive statistics and parallel coordinate plots analysis in the quantitative studies implemented during Study 1, the results obtained from other studies would address RQ 2, and independent studies would not correspond to RQ 2. The results of RQ 2 are addressed in Chapter 5, by integrating the results from a mixed methods view. In the final phase, the mixed methods approach was used to integrate the results of each study and to examine comprehensively the effectiveness of schematized feedback for students and teachers. This approach is also discussed in Chapter 5. The details of each analysis method are described in Section 3.4.

Table 3.9

Summary of the RQs and the Studies

Study	RQ (Claim in the AUA)	Analysis / *data source*
Preliminary Study (quantitative)	**RQ 1**: Reliability/consistency of using the modified ESL CP rating scale **[Claim 4]**	● Descriptive statistics ● MFRM ■ *6-level scores on writing skill criteria*
Studies 1, 3-1, and 3-2 (quantitative and qualitative)	**RQ 2**: Meaningfulness of the scores and annotation diagrams **[Claim 3]**	● Descriptive statistics ● Annotation scheme ● Parallel coordinate plots ● Thematic analysis ■ *Total scores, 6-level scores* ■ *Annotated tree diagram* ■ *Students' responses to the questionnaires*
Study 1 (quantitative)	**RQ 3 (1)**: Effectiveness of schematized feedback **[Claim 1]**	● Descriptive statistics ● Mixed-between-within MANOVA ■ *6-level scores on writing skill criteria*
Study 2 (quantitative)	**RQ 3 (2)**: Effectiveness of schematized feedback **[Claim 1]**	● Descriptive statistics ■ *Operation time efficiency of revision work referring to teacher's feedback*
Study 3 (qualitative)	**RQ 3 (3), (4)**: Effectiveness of schematized feedback **[Claim 1]**	● Annotation scheme ● Anomaly analysis ● Thematic analysis ■ *Frequency and types of coherence anomalies* ■ *Annotated tree diagram* ■ *Students' responses to the questionnaires*
Synthesis of quantitative and qualitative findings	Claim 1: Effectiveness of schematized feedback to both teachers and students	Integration of all of the above data and methods

In the pretest, each group was assigned Topic A or B randomly as the initial task to avoid an order effect, and all participants completed both tasks over two consecutive weeks. In the posttest, the students completed two consecutive weeks of revision writing based on feedback in the same order as they completed the initial task. Finally, all students worked on Topic C as the transfer task.

The questionnaires were administered with a paper-based descriptive questionnaire immediately after the rewriting task of the second topic and after the transfer task, and the responses were collected on the spot. The timing of implementation of the questionnaire along with (re)writing assignment and provision of teacher's written feedback is shown in Table 3.10. The questionnaire is provided in Appendix D

Table 3.10

The Itinerary of (Re)Writing Assignments and Questionnaire Implementation Along With the Timing of Feedback in the Classroom

Week 8-9	Week 10	Week 11		Week 12	Week 13	
Pre-tasks on two topics (30 min.)	1st FB provided + Post-task of rewriting work 1 (no time limit)	2nd FB provided + Post-task of rewriting work 2 (no time limit)	**QSA on Q. 1 and Q.2 (approx. 30 min.)**	Score results and comments returned on two revised writings	Retention task (30 min.)	**QSA on Q.2 (approx. 30 min.)**
Task (Topic) assigned						
A=>B B=>A	Revise A Revise B	Revise B Revise A			C	

Note. FB = feedback, QSA = questionnaire

The students also received overall comments and analytic scores on the web-based class platform. Then they were asked to revise their writing accordingly. The 50 students worked on two topics (Topics A and B), so the teacher's feedback was returned sequentially for a total of 100 writing samples.

As for Questionnaires, responses were obtained from a total of 45 students in the study (22 in the control group and 23 in the intervention group). For Question 1, a total of 178 responses on five criteria from 36 participants (18 from the control group and 18 from the intervention group) were collected. The number of students who responded to Question 1 is less than the total number of 45 because some students skipped Question 1 and answered only Question 2. Question 2 received 45 comments; each comment was segmented, resulting in a final total of 140 text units subjected to thematic analysis.

The data for this study was collected from weeks 8–13 of a 15-week course in the second semester of 2021. In the previous semester, the students had learned how to write a stand-alone paragraph of an opinion statement using counterargument and rebuttal as well as various other types of paragraphs. In both semesters, the instructor of the target classes was the author alone, and there was no student change.

Just prior to the research assignment, the students had reviewed argumentation during weeks 7 and 8 of the class. Refer to Appendices E and F for a sample of the instruction materials used in the class. Appendix E is an illustration of argumentation with counterargument and rebuttal of simplified topic of "Which is a better pet: a dog or a cat?" Appendix F includes text-based examples with the topic of "Which is better for movies: dubbed or subtitled?" The teaching materials were designed to take into account the differences in the two groups' forms of feedback. Materials in both forms were used for both the control and intervention groups, and the same instruction was used for both groups. The itinerary for the assignment implementation in the classroom is shown in Table 3.10.

3.4 Data Analysis

3.4.1 Quantitative Analysis (Preliminary Study, Study 1, and Study 2)

3.4.1.1 Conversion of Analytic Scores to 6-Point Scores. Before proceeding to the main analysis, a relation between the prime scores for each criterion and the scores at the six levels (see Table 3.3) was explored to examine the degree of correspondence of each pair with the approximate value curves of linear predictions on a scatterplot of scores and six levels. The R^2 values for each criterion are as follows: content, 0.9169; organization, 0.9137; language use, 0.9359; and vocabulary, 0.9293. Regarding mechanics, the raw scores and levels are identical ($R^2 = 1$) because the raw score (1–5) was used as a level without conversion. In view of this, analysis in the section with a many-facet Rasch measurement (MFRM) and subsequent sections was based on the 6-point score. However, a prime score out of 100 is sometimes used along with the 6-point score when necessary.

3.4.1.2 MFRM (Preliminary Study). In order to address RQ 1, MFRM (Linacre, 1989; Linacre & Wright, 1993; McNamara, 1996) was applied to investigate the reliability and appropriateness of using the modified ESL CP rating scale in terms of students' level, rating consistency and severity, task difficulty, and level discrimination of the writing skills criteria. The scores of the two raters were used as two independent ratings.

An MFRM is a measurement model that is suitable to simultaneously analyze multiple variables that may affect assessment outcomes. An MFRM can estimate ability values and difficulty variables, see how well the data fits the model, and examine the appropriateness of the assessment items. The procedure uses the natural logarithm of the raw values to obtain values on a logit scale; these transformed values are used to estimate the features of the item and the ability of the examinee. For each element of each facet, an

MFRM analysis provides fit indices to show the degree to which observed ratings match the expected ratings generated by the model. The partial credit model (PCM) estimates threshold parameters for each item separately, allowing each item to have a unique rating scale structure (Eckes, 2015, p. 28). The present study applied the PCM (Masters, 1982, 2010) because it is assumed the relative difficulty between criteria is expected to vary across writing skill criteria. The criterion-related four-facet PCM, which corresponds to the present study design, states that test-takers, tasks, rating criteria, and raters can be specified with Equation 3.1 (Barkaoui, 2013, p. 1304; Eckes, 2015, p. 128):

$$log\ (P_{nmijk} / P_{nmijk-1}) = \mathrm{B}_n - \mathrm{D}_m - \mathrm{E}_i - \mathrm{C}_j - \mathrm{F}_{ik}, \qquad (3.1)$$

where

P_{nmijk} = the probability of test-taker *n* achieving task *m*, for criterion *i* by rater *j* a score *k*;

$P_{nmijk-1}$ = the probability of test-taker *n* achieving task *m*, for criterion *i* by rater *j* a score *k-1*;

$log\ (P_{nmijk} / P_{nmijk-1})$ = the log odds of achieving a score k, given the task, criterion, and rater, versus the probability of being rated k-1;

B_n = the ability (B) of test-test taker *n*;

D_m = the difficulty (D) of task *m*;

E_i = the difficulty of (E) of criterion *i*;

C_j = the severity (C) of rater *j*;

F_{ik} = the difficulty of receiving a rating of *k* relative to *k-1* on criterion *i*.

The data for the MFRM analysis consisted of 1,800 valid ratings assigned by six raters and 45 students (22 from the control group and 23 from the intervention group) on a total of five tasks, and with four rating criteria. However, for raters, in practice, one writing was scored by a combination of two different raters, so it is not *two raters* but *two ratings*. Table 3.11 shows the number of valid ratings included in the MFRM analysis in FACETS.

Table 3.11

The Number of Valid Ratings Included in the MFRM analysis in FACETS

Facet	Students	Raters	Tasks	Rating criteria	Number of valid ratings
Breakdown	22 + 23	Six raters split into three groups of two	Pre- and posttests for Tasks A and B, plus Transfer C task	Content, organization, language use, and vocabulary	45 × 2 × 5 × 4
Number of elements	45	2	5	4	1,800

Regarding the sample size for an MFRM, Barkaoui (2013) cited a personal communication with Linacre in 2012 that a reasonable FACETS analysis would contain at least 900 data points with at least two elements in each facet. The present study does not have a large sample size, but there are a total of 1800 data points, which exceeds Linacre's recommendation of 900 data points. Therefore, it was determined that it would hold up to analysis in FACETS. Regarding assumption checks for conducting the MFRM analysis, unidimensionality and a global model fit were tested (Brentani & Golia, 2007; Eckes, 2015).

3.4.1.3 Parallel Coordinate Plots. This analysis was conducted to investigate the characteristics of the students' overall writing performance to address RQ 2. Before comparing mean scores, a parallel coordinate plot was created to capture changes in the learners' scores over time at the individual level using total scores. Tests using group mean scores are useful to objectively identify overall trends as a group, but such analyses sometimes have the potential weakness of neglecting individual information. This in turn requires more careful observation of an independent student's score data because individual evaluations and changes are also important in classroom assessment. By referring to Adamson and Bunting (2005), Larson-Hall (2010) states that a parallel coordinate plot (also called a profile plot) can provide viewers with many more points of data than the means plot, which enables offering a general impression of the trends of individuals. The reason for using the 100-point prime score rather than the 6-point score in this analysis is that changes at the individual level are not intended to be generalizable, so we do not want to miss any small changes with a more careful observation than is intended. This graphic reveals the existence of individuals who move differently from the average by visualizing individual movements that are difficult to see with just the means. This approach brings formative assessment in the classroom closer to reality.

3.4.1.4 Mixed-Between-Within Multivariate Analysis of Variance. A mixed-between-within multivariate analysis of variance (MANOVA), hereafter called mixed MANOVA, was conducted to address RQ 3(1). This mixed-design statistical analysis was applied because there is a between-groups independent variable (two different groups based on feedback type: conventional versus graphic) and a within-groups independent variable (three writing occasions in a repeated design). Moreover, there are multiple dependent variables (four writing skill criteria: content, organization, language use, and vocabulary).

Prior to conducting the mixed MANOVA, preliminary checks were conducted on the assumptions of normality and homogeneity of variance-covariance matrices (based on Box's test of equality of covariance matrices), and the linearity between the three writing occasions for each of the four criteria and between the groups (based on regression analysis and scatter plots). Furthermore, the values of bivariate correlations were examined to test the MANOVA assumption that the dependent variables are correlated with each other in the moderate range (Meyer et al., 2006).

MANOVA is advantageous because it allows for an analysis that considers correlations in the dependent variable. This method is particularly appropriate in the case of this study, where the four dependent variables of writing ability are presumed to be related to each other.

3.4.1.5 Time on Task in Each Group (Study 2). To address RQ 3 (2), the effects of schematized feedback on the time required for the process writing revision work were investigated. For this endeavor, the operation time on task and the total number of words were compared between the control and intervention groups. All work was produced using word processing software on a computer and submitted to the university web portal, where the time of submission was recorded. The work time was calculated from the submission time. The number of words written per minute was also determined to evaluate efficiency.

3.4.2 Qualitative Analysis (Study 3)

3.4.2.1 Coding Scheme Based on Rhetorical and Ideational Anomaly Detection. This analysis was conducted to address RQs 3 (3) and (4), to determine the effectiveness of schematized feedback. First, to proceed with the qualitative analysis of the organization of the students' writing, a taxonomy and definitions of anomalies, which can be used as one of the indicators to evaluate organization, was proposed. This information was accompanied by the descriptors of the modified ESL CP (Jacobs et al., 1981) used for the quantitative evaluation in the study, and classification of rhetorical and ideational rater judgment strategies in assessing EFL writing proposed by Barkaoui (2007) and Cumming et al. (2002). The anomalies identified in the writing samples were coded based on this coding scheme by two annotators. The inter-annotator agreement on the location of the anomalous units in discourse and that for the anomaly types based on the newly developed taxonomy was calculated. Of the 100 identified anomalous text units, the degree of agreement at the outset was 82% for the anomaly location and 71% for the 11 newly developed classifications.

The proposed taxonomy attempted to categorize rhetorical and ideational coherence, based on the identified anomalies. Neither Barkaoui (2007) nor Cumming et al. (2002) necessarily defined all the strategies clearly, nor did they distinguish between rhetorical and ideational foci. In their studies, these two were treated as a single category. However, Cumming et al. (2002) provided some representative comments by the raters in their appendix. In the present study, therefore, while referring to the descriptors in the ESL CP and anomalies identified in the samples, nine strategies that were originally considered as a single entity were tentatively divided into four rhetorical and five ideational strategies. In addition, two language focus strategies were included in the coding scheme because they were among the anomalies identified in the annotation process. On the other hand, the other seven language focus strategies[11] introduced by Cumming et al. (2002) are not included in the taxonomy here because the problems in them are not of the nature of what is judged by tagging in the annotation tool. These classifications were used in a later qualitative analysis of students' writing samples.

3.4.2.2 Thematic Analysis on Students' Responses to the Questionnaire. The students' comments to questionnaires (see Section 3.2.7) were investigated by thematic analysis following the procedure proposed by Takagi (2021). The comments were to question

[11] The seven other language focus strategies that are not included in the analysis are: consider gravity of error, consider error frequency, assess fluency, consider lexis, consider syntax or morphology, consider spelling or punctuation, and rate language overall.

1 regarding the revisions provided during the rewriting process in the posttest and the transfer tasks and to question 2 on the teacher feedback.

Question 1 was given immediately after the rewriting task of the second topic (either Topic A or B) in week 11. The participants were asked to describe what they had revised in the five analytic criteria—content, organization, vocabulary, language use, and mechanics—and the comments to the teacher feedback provided prior to the revising task. In this section, the responses regarding the modifications were analyzed first. A standard thematic analysis was conducted on the data obtained from 36 participants (18 from the control group and 18 from the intervention group), which is less than the total number of the participants because some of the students responded to the questions regarding the feedback but not to the revisions.

Question 2 was given on two separate occasions: immediately after the second revision task along with question 1 in week 11 and after the Transfer C task in week 13. The question was given on two separate occasions to give the students an immediate impression of their own writing as they revise, referring back to the feedback as they revise, and to have them look back on the feedback again from a bird's eye view after working on a new assignment on a new topic in the following week. However, because of the short interval between these two occasions, both responses were combined for analysis and tabulation. There were 45 respondents, 22 in the control group and 23 in the intervention group. Each of the comments were segmented into units for analyses. A total of 140 units were identified, 71 from the control group and 69 from the intervention group. The resulting units were coded in an exploratory manner.

This study adopted the procedure proposed by Takagi (2021), which is based on the procedures of Braun and Clarke (2006), Creswell and Creswell Báez (2021), and Flick (2014). The procedure is shown below:

1. Decide on an RQ

2. Familiarize yourself with your data

3. Divide text into segments of information

4. Code the data

5. Evaluate and modify the coding

6. Develop themes or categories

7. Interpret the meaning of the themes of categories

8. Report the interpretations or findings

Finally, the themes obtained through thematic analyses were summarized into an illustration of a storyline to explains a series of students' behaviors in the process of rewriting (Creswell & Creswell Báez 2021).

Chapter 4

Results

4.1 Quantitative Research Results

4.1.1 Quantitative Study Results for Study 1: RQ 1, RQ2, and RQ3(1)

This section summarizes the analysis of 225 writing samples produced by 45 students using statistical methods to address RQ 1, the preliminary study that examined the reliability of the rating scale and the qualitative part of the two RQs: RQ 2, which investigates the meaningfulness of interpretation of the students' overall writing performance, and RQ 3(1), which explores the effectiveness schematized teacher feedback. For the sake of convenience, the first/initial draft are referred to as the pre-task and the second draft revising the initial draft is referred to as the post-task. For example, the first draft of Topic A is called the Pre A task and the second draft of Topic B is called the Post B task. Finally, the transfer task on Topic C is Transfer C.

4.1.1.1 RQ 1: Analysis of the Appropriateness of the Rating Scale. For RQ 1, an MFRM (Linacre, 1989; Linacre & Wright, 1993; McNamara, 1996) approach was applied with the FACET program.[12] Because there were < 2,000 observations in this study, Minifac Version No. 3.81.1[13] was used for the analysis. The aim was to investigate the appropriateness of the rating scale with the following four facets: examinees; raters; tasks; and the content, organization, language use, and vocabulary rating criteria. This examination

[12] https://www.winsteps.com/a/Winsteps-Manual.pdf

[13] https://www.winsteps.com/minifac.htm

of reliability was conducted because the 6-point rating scales and their descriptors were developed for this study from the four-level ordinal scale of the ESL CP (Jacobs et al., 1981). Subsequent quantitative analyses were mainly conducted with the 6-point scores of the rating scale.

4.1.1.1.1 Assumption Checks for the MFRM Analysis. Preliminary assumption checks for the MFRM analysis showed that the unidimensionality of data and global model fit were acceptable. First, regarding the unidimensionality, the variance explained by Rasch measures was 43.56%. The data derived from the principal component analysis (PCA) of standardized residuals (Chou & Wang, 2010; Linacre, 1998, 2014; Smith 2002) showed that the raw score-score variance of observations was 1.468 (100.00%), the variance explained by Rasch measures was 0.639 (43.56%), and the variance of residuals was 0.828 (56.44%). According to Engelhard (2013, p. 185), unidimensionality is satisfied if the variance explained by Rasch measures is ≥ 20%. Therefore, the unidimensionality of the present data was satisfied.

Second, with regard to the global model fit, a total of 1,800 responses were analyzed to estimate parameter values. Of these, 69 responses (or 3.83%) were associated with absolute standardized residuals ≥ 2, and 8 responses (or 0.44%) were associated with absolute standardized residuals ≥ 3. A model fit is satisfactory when ≤ 5% of absolute standardized residuals are ≥ 2, and ≤ 1% of absolute standardized residuals are ≥ 3 (Linacre, 2014). Therefore, these results suggest a satisfactory model fit. The log-likelihood chi-square value was 4575.1470 (df = 1732, p < .00). According to Eckes (2015, p. 69), if the results are statistically significant, which applies to this case, the data cannot be said to fit the model in Rasch analysis. However, he also mentions that this could happen for nearly any set of empirical observations. Thus, the data of the present study can be interpreted to fit the Rasch analysis sufficiently well, considering that the other preliminary assumption checks were satisfied.

4.1.1.1.2 Examining Fit Statistics for MFRM. To determine the degree to which the observed ratings matched the expected ratings generated by the model, fit indices were examined for the facets included in the model: students, ratings, tasks, and the four writing skills criteria. They refer to the extent to which a given measure corresponds to Rasch model expectations (Linacre, 2004). The mean-square infit statistic (MS_w) and the mean-square outfit statistic (MS_U) are residual-based indices of how the data fit a model. Standardized statistics (*Z* std) obtained by standardizing *MS* statistics are also provided in the measurement report.

There are some guidelines concerning the range of acceptable values for mean-square fit statistics in the literature. Linacre (2002, 2014) suggests that a range of 0.5 to 1.5 for both MS_w and MS_U is "productive for measurement" and all others are misfit. In addition, a standardized value between -2 and +2 is regarded as indicative of a useful fit (Bond & Fox, 2007, p. 43). Although stricter standards exist for high-stakes tests, the guideline suggested above was used for this study, given that the test was for low-stakes formative classroom assessment with a small sample size.

Table 4.1 shows the measurement results about the three facets (writing skill criteria, rating, and task). Here, three occasions (the first-draft is referred to as the pre-task and the second-draft is referred to as the post-task) and two task combinations were analyzed as separate tasks, for a total of five tasks (Pre A, Pre B, Post A, Post B, and Transfer C). For the student facet, a simplified table summarizing the percentage of model fits is presented due to space limitations (Table 4.2).

Table 4.1

Measurement Report Regarding the Three Facets: Writing Skills Criteria, Ratings, and Tasks

Facet	Observed Average	Measure logit	Mode *SE*	Inft *M* Sq	*Z* std	Outfit *M* Sq	*Z* Std
Criteria							
Vocabulary	3.66	.64	.05	.92	-1.20	.65	-1.10
Content	3.65	-.11	.05	1.04	.50	1.04	.60
Organization	3.65	-.21	.05	1.11	1.60	1.12	1.80
Language use	3.47	-.32	.05	.95	-.70	.97	-.40
Mean (count: 4)	3.61	.00	.05	1.01	.10	1.01	.20
SD (population)	.08	.38	.00	.07	1.10	.07	1.10
SD (sample)	.09	.44	.00	.08	1.30	.08	1.30
Rating							
Rating 1	3.60	.00	.04	1.09	1.90	1.09	1.90
Rating 2	3.61	.00	.04	.92	-1.70	.94	-1.30
Mean (count: 2)	3.61	.00	.04	1.01	.10	1.01	.30
SD (population)	.00	.00	.00	.08	1.80	.08	1.70
SD (sample)	.01	.01	.00	.12	2.60	.11	2.30
Task							
Pre B	2.85	.94	.06	1.20	2.40	1.15	1.80
Pre A	3.28	.37	.06	.91	-1.20	.94	-.80
Transfer C	.07	-.13	.06	.98	-.20	.98	-.20
Post B	3.95	-.42	.06	1.25	3.10	1.27	3.60
Post A	4.26	-.76	.06	.72	-4.40	.74	-4.10
Mean (count: 5)	3.61	.00	.06	1.01	.00	1.01	.00
SD (population)	.05	.60	.00	.19	2.80	.18	2.60
SD (sample)	.55	.67	.00	.21	3.10	.20	2.90

Table 4.2

Model Fit (%) for the Student Facet

Facet	Overfit	Fit	Underfit
Students	2.22(1/45)	93.33 (42/45)	4.44 (2/45)

In Table 4.1, the MS fit statistics are between 0.5 and 1.5, although there are a few values that deviate from the absolute *Z st*d value of 2. The MS values are considered "productive for measurement" or as indicative of "useful fit" (Eckes, 2015, p. 80; Linacre, 2003), while standardized fit statistics test are often used for the purposes of significance testing. Therefore, the rating scale in this study was considered to fit the model in terms of the three facets. In Table 4.2, 93.33% of students in the student facet fit the model while 2.22% are overfit and 4.44% are underfit (misfit). Overfit items suggest too little variation or too determined of a response pattern, whereas underfit (or misfit) indicates a response pattern that is too haphazard or too much variation (Bond & Fox, 2007, p. 240). Overfit items are somewhat harmless, but misfit items should be handled with care and the cause of the occurrence should be explored. After reviewing the writing samples of the two students who fell into misfit, both students scored high on most tasks and criteria but scored exceptionally low on content and organization. They seemed unable to successfully develop the idea in certain topics. It is possible that some students were not able to proceed successfully with the idea, a phenomenon that can occur in opinion paragraph writing. Therefore, these students were kept for further analysis.

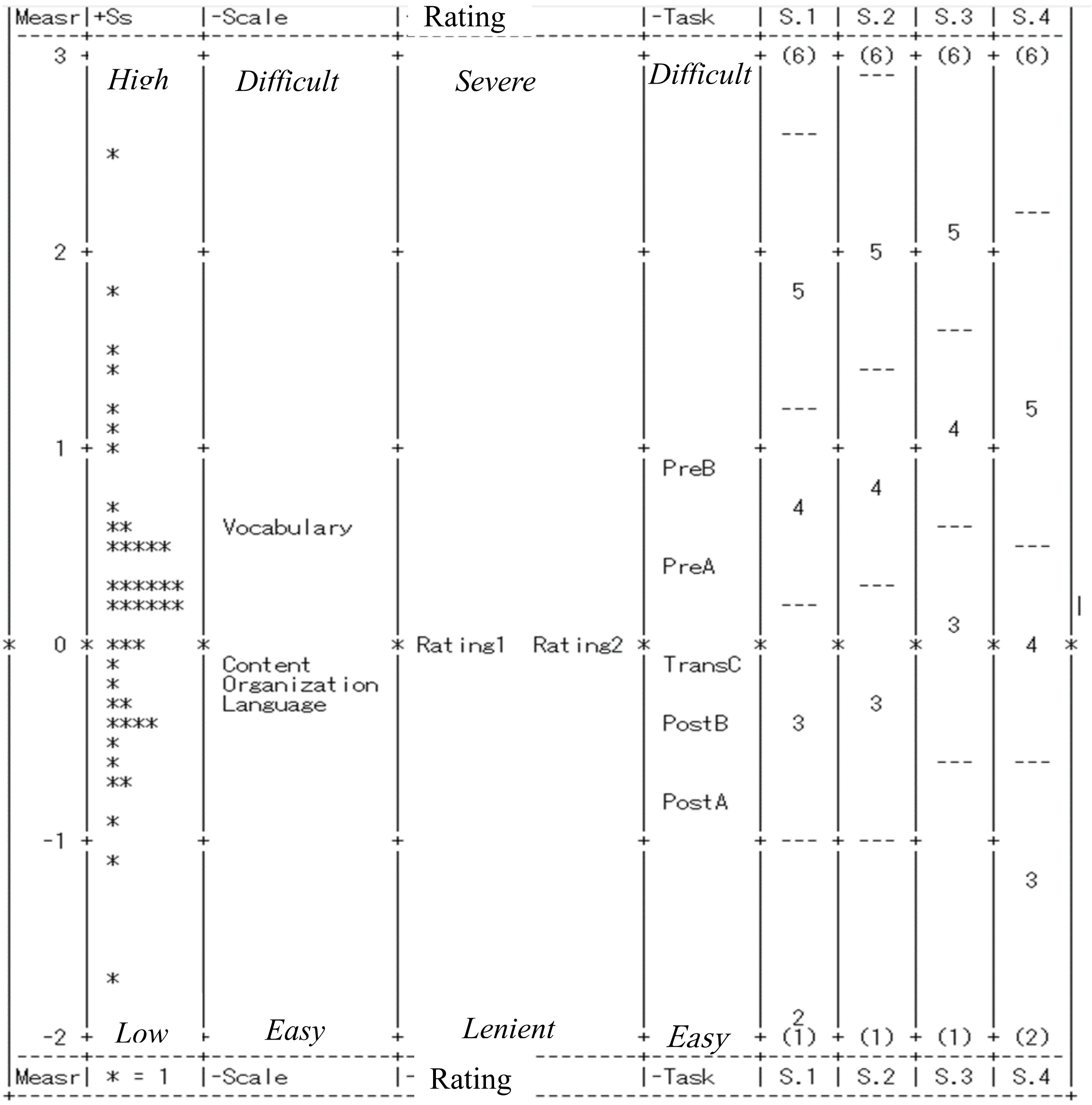
Based on the results above, the data fit the model well enough, and the model fit indices were acceptable and without problems.

4.1.1.1.3 The Wright Map for the Joint Calibration of the Four Facets. Before presenting the results of the analysis, it should be noted that mechanics was not included in the MFRM analysis because its level structure is quite different from the other criteria and unique in its nature and rating approach.

Figure 4.1 is a graphical display showing the joint calibration of test-takers (students), scales (criteria), raters (ratings), tasks, and ability level scale based on the model presented in Equation 3.1. The overall view of the measurement results shows relations between and within facets, and category thresholds can be captured by the vertical rulers at glance. Moreover, a careful observation of the vertical map provides guidance on how to interpret the assessment results, the appropriateness of the assessment design, as well as information on rating effectiveness or scale quality along with data-model fit statistics computed on a category basis. An explanation of the map is given after the figure. The description follows Eckes (2015, pp. 58–60).

Figure 4.1

Wright Map From the MFRM Analysis



```
|Measr|+Ss     |-Scale        |-Rating                |-Task    | S.1 | S.2 | S.3 | S.4 |
|-----+--------+--------------+-----------------------+---------+-----+-----+-----+-----|
|   3 +        +              +                       +         + (6) + (6) + (6) + (6) |
|     |High    | Difficult    |        Severe         |Difficult|     | --- |     |     |
|     |        |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     |     |     |
|     |        |              |                       |         | --- |     |     |     |
|     | *      |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     |     | --- |
|     |        |              |                       |         |     |     |  5  |     |
|   2 +        +              +                       +         +     +  5  +     +     |
|     |        |              |                       |         |     |     |     |     |
|     | *      |              |                       |         |  5  |     |     |     |
|     |        |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     | --- |     |
|     | *      |              |                       |         |     |     |     |     |
|     | *      |              |                       |         |     | --- |     |     |
|     |        |              |                       |         |     |     |     |     |
|     | *      |              |                       |         | --- |     |     |  5  |
|     | *      |              |                       |         |     |     |  4  |     |
|   1 + *      +              +                       +         +     +     +     +     |
|     |        |              |                       | PreB    |     |     |     |     |
|     |        |              |                       |         |     |  4  |     |     |
|     | *      |              |                       |         |  4  |     |     |     |
|     | **     | Vocabulary   |                       |         |     |     | --- |     |
|     | *****  |              |                       |         |     |     |     | --- |
|     |        |              |                       | PreA    |     |     |     |     |
|     | ****** |              |                       |         |     | --- |     |     |
|     | ****** |              |                       |         | --- |     |     |     |
|     |        |              |                       |         |     |     |  3  |     |
*   0 * ***    *              * Rating1  Rating2      *         *     *     *     *  4  *
|     | *      | Content      |                       | TransC  |     |     |     |     |
|     | *      | Organization |                       |         |     |     |     |     |
|     | **     | Language     |                       |         |     |  3  |     |     |
|     | ****   |              |                       | PostB   |  3  |     |     |     |
|     | *      |              |                       |         |     |     |     |     |
|     | *      |              |                       |         |     |     | --- | --- |
|     | **     |              |                       |         |     |     |     |     |
|     |        |              |                       | PostA   |     |     |     |     |
|     | *      |              |                       |         |     |     |     |     |
|  -1 +        +              +                       +         + --- + --- +     +     |
|     | *      |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     |     |  3  |
|     |        |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     |     |     |
|     | *      |              |                       |         |     |     |     |     |
|     |        |              |                       |         |     |     |     |     |
|     |        |              |                       |         |  2  |     |     |     |
|  -2 + Low    + Easy         +        Lenient        + Easy    + (1) + (1) + (1) + (2) |
|-----+--------+--------------+-----------------------+---------+-----+-----+-----+-----|
|Measr| * = 1  |-Scale        |-Rating                |-Task    | S.1 | S.2 | S.3 | S.4 |
+---------------------------------------------------------------------------------------+
```

Note. Each star in the second column represents a student. The horizontal dashed lines in the four columns on the right indicate the level threshold measures. S.1, content; S.2, organization; S.3, language; S.4, vocabulary.

In the Wright map all the measures of students, criteria, raters, tasks, and ability level scales are positioned vertically on the measurement scale (in logit) in the leftmost column.

The following describes what the map indicates for each facet.

The second column ("Ss," i.e., students) displays the estimates of the students' ability level. In the column, those with higher scores are placed higher and those with lower scores are placed lower. The peaks of the distribution of students appear between 0 and +1 logit, indicating that the writing assessment in this study was appropriate for the students' ability levels. Additionally, the overall distribution of students appears to be generally normal, which also indicates the appropriateness of the assessment for the target students.

The third column ("Scale") compares the four scoring criteria of content, organization, language use, and vocabulary in terms of their relative difficulties. Those located higher in the column were more difficult than those located lower in the column. Vocabulary was the most difficult, whereas the other three appeared less difficult. The difference in difficulty between the latter three is not great at all, but the order from the most to least difficult is content, organization, and language. It should be noted here that none of the items are extremely difficult or easy, as they all fall between -1 and +1 logit.

The fourth column ("Ratings") compares ratings in terms of the level of severity in evaluating paragraph writing. The fact that the two ratings are aligned side by side, and furthermore, that they are located at zero logit, suggests that there was very little discrepancy between the ratings, and that they were appropriately handled without being too severe or too lenient.

The fifth column ("Task") compares the five tasks, while treating pre- and post-tasks as independent assignments, in terms of their relative difficulties. More difficult tasks appear higher in the column, while less difficult tasks appear lower in the column. Thus, the mean score for the higher-positioned tasks will be lower because they are more difficult to score. The display in this column shows that the pre-tasks were harder and the post-tasks were easier, while Transfer C was somewhere in between in terms of the difficulty level. More importantly, the Pre and Post B tasks were consistently more difficult than the Pre and Post A tasks. This result is consistent with the results for the test of task difficulty presented in Section 4.1.1.1"

The last four columns represent the 6-point rating scales on the logit scale, with each column corresponding to a respective criterion. As described in Chapter 3, the present study applied the PCM (Masters, 1982, 2010). In the PCM, each rating scale for each criterion is modeled to have its own category structure. The lowest scale category and the highest scale category are shown in parentheses. The horizontal dashed lines in the four columns are positioned at the category thresholds.

Table 4.3 provides the corresponding specific data for Figure 4.1. It shows the measurement reports of the four facets: students, rating criteria, ratings, and tasks.

Table 4.3

Measurement Report of the Four Facets Obtained with MFRM

Facet	*M* (measure)	*SD* (measure)	Min, Max	Range	Separation or strata index	Separation reliability
Students	.20	.76	-1.68, 2.48	4.16	5.9 (strata)	.95
Criteria	.00	.38	-.32, .64	.96	7.2 (separation)	.98
Ratings	.00	.00	.00, .00	.00	.00 (separation)	.00
Tasks	.00	.60	-.76, .94	1.7	10.19 (separation)	.99

Most of the information in Table 4.3 aligns with the information presented in the vertical map discussion. Furthermore, there is some additional information on “separation or strata” and “separation reliability.” According to Linacre (2012, pp. 304–305), the choice between the separation or strata index depends on the characteristics of the distribution. The strata index was chosen for the student facet because the distribution was assumed to be caused by differences in ability, and thus the distribution was not necessarily normal. For the other facets, the separation index was chosen by examining the actual values. The homogeneity in the rating facet could be attributed to the fact that the five raters’ scores were integrated as rating 1, reducing the heterogeneity of ratings, but smaller differences in severity levels between reliability is, high inter-rater reliability, is preferable for the rating data. For the criteria and tasks facets, the separation values were about 7 and 10, respectively. These values are greater than the actual number of the items included in the analysis. This result suggests that the spread of the criterion difficulty measures was greater than the precision of those measures. As Eckes (2015) remarks, “Generally speaking, high separation is caused by a large number of observations available for each element in the facet and/or a large “true” standard deviation of the measures for each element” (p. 65). There are cases where it is preferable to be compatible for all tasks, but in this study, Topics A and B are at different levels of the original English test, and it is expected that there will be differences between the Topics, so this phenomenon is assumed to be fine. The statistical difference between Topics A and B are presented in Section 4.1.1.2.

4.1.1.1.4 Functioning of the Scale. The category statistics of the score count distribution of two ratings are presented in Table 4.4. The effectiveness of the rating scale was examined by following the rating scale quality indicators and guidelines provided by Eckes (2015, p. 117) based on Linacre (1999, 2004). The following six points indicate a high scale quality: (1) the number (*N*) of responses per category ($N \geq 10$), (2) response frequency across categories (regular; uniform, unimodal, bimodal), (3) the average measures by category (monotonic increase with category), (4) the model fit of rating scale ($MS_u < 2.0$), (5) threshold order (monotonic increase), and (6) and size of threshold increase (≥ 1.4 and < 5.0 logit).

.

Table 4.4

Category Statistics of Score Count Distribution of Two Ratings

	Content			Organization			Language			Vocabulary		
Category	Total	%	Cum.%	Total	%	Cum.%	Total	%	Cum.%	Total	%	Cum.%
1	11	2%	2%	6	1%	1%	2	0%	0%	0	0%	0%
2	77	17%	20%	85	19%	20%	123	27%	28%	73	16%	16%
3	117	26%	46%	115	26%	46%	122	27%	55%	151	34%	50%
4	135	30%	76%	132	29%	75%	96	21%	76%	108	24%	74%
5	74	16%	92%	80	18%	93%	79	18%	94%	92	20%	94%
6	36	8%	100%	32	7%	100%	28	6%	100%	26	6%	100%
Total 450 ratings				450			450			450		

The threshold measures in the rating scale category calibration are given in Table 4.5. This information is useful to examine the extent to which the scale functions effectively.

Table 4.5

Rating Scale Category Calibrations for Each Criterion

Category	Content		Organization		Language use		Vocabulary	
	Threshold measure	*SE*	Threshold measure	*SE*	Threshold measure	*SE*	Threshold measure	*SE*
Level 2	-2.74	0.32	-3.37	0.42	-4.64	0.71	-	-
Level 3	-0.69	0.14	-0.48	0.14	0.04	0.12	-1.79	0.14
Level 4	0.09	0.11	0.20	0.11	0.78	0.12	-0.19	0.11
Level 5	1.35	0.13	1.34	0.13	1.22	0.13	0.15	0.13
Level 6	1.99	0.20	2.32	0.21	2.61	0.22	1.83	0.22

Note. The thresholds are Rasch–Andrich thresholds.

As Table 4.4 suggests, there are > 10 responses per category in most categories except for the three categories of the ability scale of category 1. Because it is important to set these six levels for instructional purposes, the distinction should not be eliminated, and the levels should not be collapsed. Next, the response frequencies across categories are regular and unimodal for the three criteria and slightly bimodal for the one criterion, content, which satisfies point 2 described above. As the calibration table in Table 4.5 shows, the average measures by category shows monotonic increase with no scale inversion nor disordered thresholds in either criterion, which would satisfy points 3 and 5. Point 4 regarding model fit of the rating scale was examined in Section; 4.1.1.1.2, there was no problem. Lastly, some of the sizes threshold increase do not fit the lower limit standard (≥ 1.4 logit) and do not seem to follow the suggested point 6. According to Eckes (2015), in the case that the thresholds are too close, the categories involved are less distinctive than intended. For remedial action, refinement or combining categories may be considered. However, he also warns by writing that Myford (Eckes' personal communication, 2015) is against easy action of collapsing categories, in particular with small samples of raters and/or examinee performances (Eckes, 2015, p. 121). As a result, the study met five of the six guidelines, and for the remaining one, given that the sample size was not large, this rating scale worked effectively to some degree and its use in low-stakes classroom assessments was determined to be appropriate.

To summarize the examination conducted in this section as a preliminary study prior to addressing RQs 2 and 3, it was concluded that the 6-point rating scale used in this study, which is a modified version of the ESL CP by Jacobs et al. (1981), has a certain degree of reliability for use in low-stakes assessment in the classroom.

4.1.1.2 Measuring Differences in Difficulty Between the Pre A and Pre B Tasks.

To examine the comparability of the tasks in terms of difficulty, the difficulty level of Topics A and B was evaluated by comparing the scoring results of the Pre A and Pre B tasks. Table 4.6 presents the descriptive statistics of the Pre A and Pre B tasks. The mean score of the Pre A task is higher than that of Pre B, which indicates the tendency that the Pre A task was less difficult than the Pre B task. A *t*-test revealed that this difference was significant: t (44) = 3.993, *p* = .0001, *r* = .52. Hence, Topic A was significantly less difficult than Topic B for the students in this study.

Table 4.6

Descriptive Statistics for the Modified ESL CP Writing Evaluation for the Pre-A and Pre-B Tasks

	M	*SD*
Pre-A	66.2	9.1
Pre-B	60.4	9.0

Note. The total score is 100 (*n* = 45).

4.1.1.3 Interpretation of the Students' Overall Performance in Terms of the Composite Score. In Section 4.1.1, four facets including each writing skill criterion in the rating scale were analyzed with the 6-point scores to examine reliability using an MFRM analysis. Likewise, the effect of schematized feedback is examined in Section 4.1.2, also using the 6-point scores, with mixed MANOVA.

This section identifies and reports the profiles of the overall writing features of students in response to RQ 2 with the 100-point composite score. This score was used to emphasize the profiling element of capturing more detailed characteristics. They were analyzed with descriptive statistics and parallel coordinate plots, which can capture the three writing tasks at the individual level.

4.1.1.3.1 Descriptive Statistics over Three Writing Occasions with Total Scores. Table 4.7 shows the descriptive statistics of the average scores of the Pre A, Pre B, Post A, Post B, Transfer C tasks. The boxplots are presented in Figure 4.2.

Table 4.7

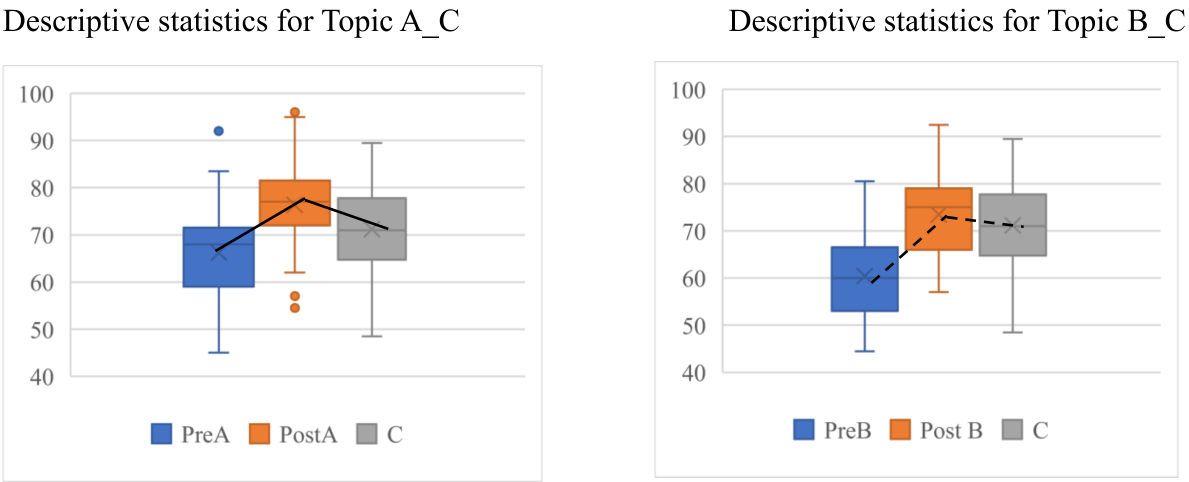
Descriptive Statistics for Modified ESL CP Writing Evaluation for Topics A, B, and C At the Three Writing Occasions

	M	*SD*
Pre A	66.2	9.1
Pre B	60.4	9
Post A	76.3	8.8
Post B	73.4	8.5
Transfer C	71.2	8.8

Note. The total score is 100 ($n = 45$).

Figure 4.2

Descriptive Statistics at Three Occasions for Topic A_C and B_C



Descriptive statistics for Topic A_C

Descriptive statistics for Topic B_C

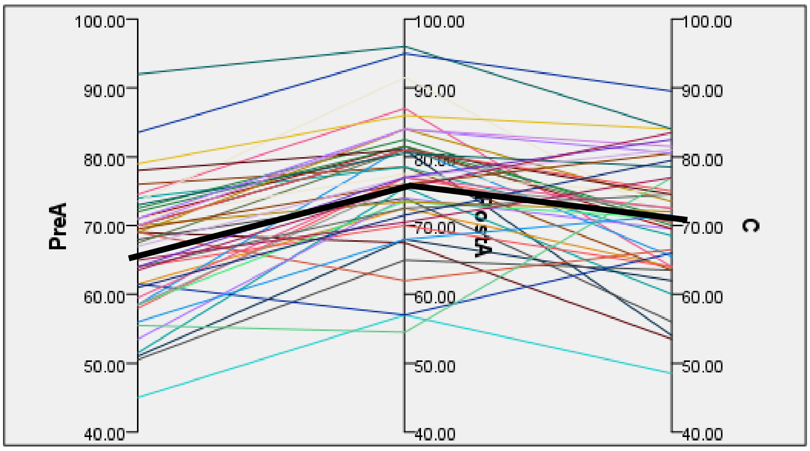
The significant difference in initial scores between Topic A and B was confirmed in Section 4.1.1.2, but in both tasks, scores rose considerably in the post-task and dropped to a certain degree in the transfer task, regardless of the difficulty of the task. In the left figure, the Transfer C task drops to the same level as the Pre A task, but the right figure shows that it does not drop as much as the Pre B task. The *SD* for all tasks remained between 8.8 and 9.0 except for the Post B task, which was 8.5. The *SD* is 8.8–9.0 for all tasks except for the Post B task (8.5). It is interesting to note that the variation is a little smaller in the Post B task, which is the more difficult one, although this may be due to the lower Pre A task score. These data indicate the compatibility of the tasks, and the trends in these scores provided meaningful information about the test scores across tasks and across occasions. It can be concluded that the information was provided in a way that the stakeholders could interpret and appreciate.

4.1.1.3.2 Parallel Coordinate Plots with Total Scores over Three Tasks and Between Groups. A parallel coordinate plot is a nice graphic display that "takes the place of the means plot and contains many more points of data" (Larson-Hall, 2010, p. 331). This plot, also known as a profile plot, offers information on the overall trend as well as the data, providing insight that cannot be obtained from averages alone. In that sense, it is an excellent way to present assessment results because it is a small classroom where individual-level data are considered to have weight.

Figures 4.3 and 4.4 show the parallel coordinate plots for students' total scores at the pretest, posttest, and retention condition. These plots provide observations that are closer to the actual situation by revealing individual changes over time as complementary information, which cannot be obtained from the overall averages alone. Figure 4.3 is the trace of the Pre A, Post A, and Transfer C tasks, and Figure 4.4 is the trace of the Pre B, Post B, and Transfer C tasks. Both figures also contain a bold line that indicates the mean.

Figure 4.3

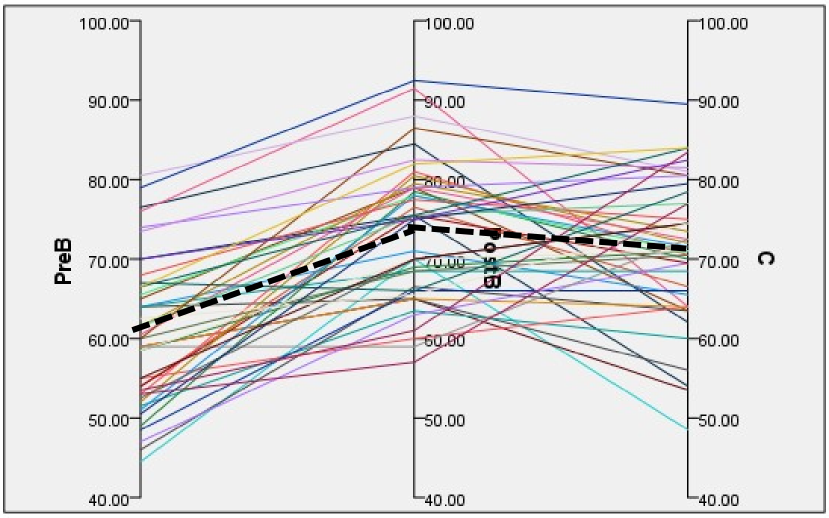
Parallel Coordinate Plots for the Pre A, Post A, and Transfer C Tasks



Note. The bold line represents the population average (*n* = 45).

Figure 4.4

Parallel Coordinate Plots for the Pre B, Post B, and Transfer C Tasks



Note. The bold dashed line represents the population average (*n* = 45).

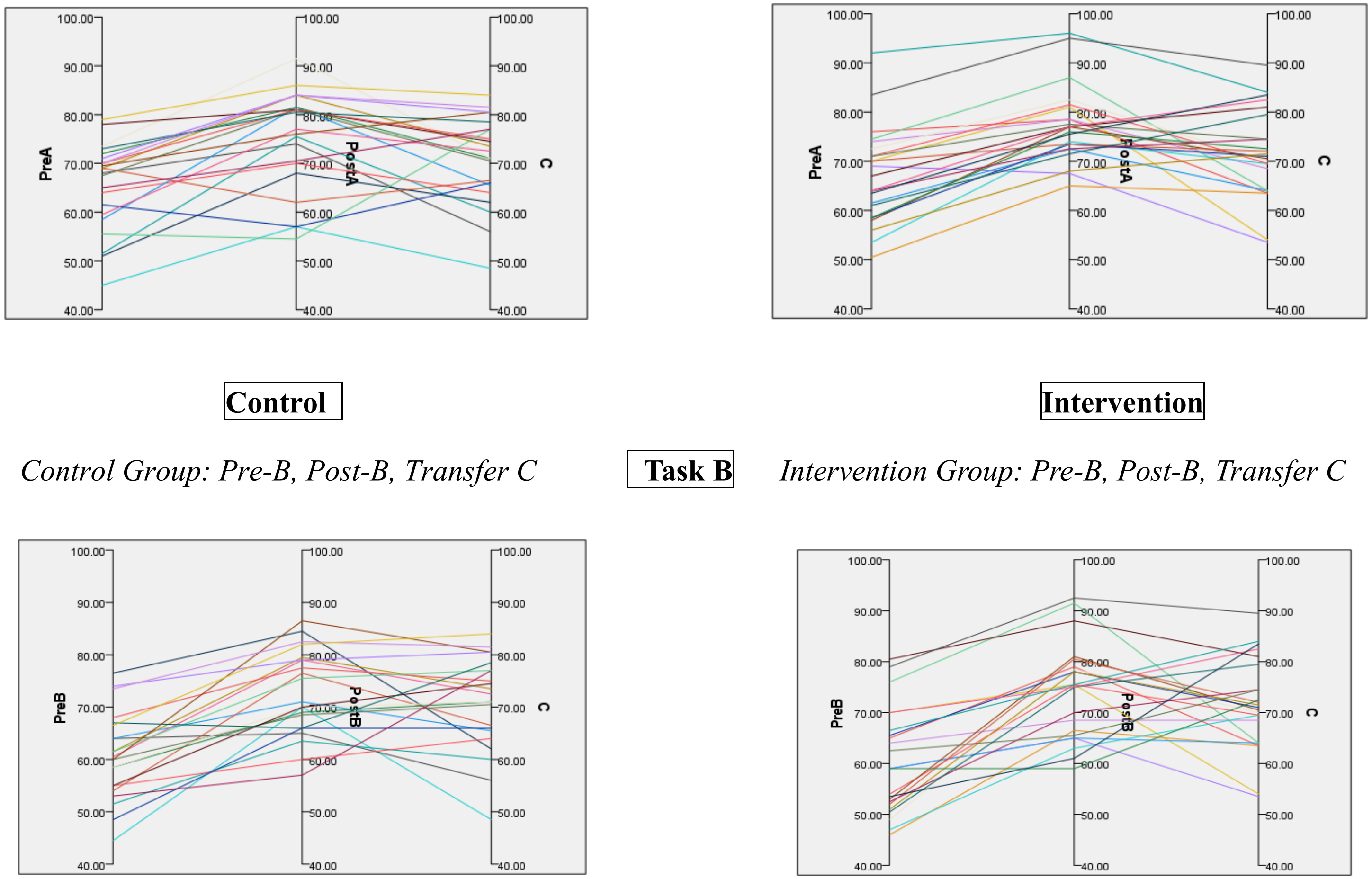
Comparing the Topic A_C and Topic B_C panels, the overall behavior appears to be similar, but a closer observation reveals that there was a general tendency for the Topic B_C group to show a linear increase in scores, starting from a lower score. Moreover, while some individuals in both groups scored significantly lower on the transfer test than on the posttest, some students who scored relatively low on the pretest continued to increase their scores, and some ultimately scored higher on the transfer test. This seems to be the case for Topic B, which is considered to be a more difficult than Topic A.

The more difficult Topic B shows a somewhat more uniform or pronounced posttest upward trend than the less difficult Topic A, although the overall mean changes for different task difficulty levels show similar trends. Furthermore, in Topic A, some of the participants' scores did not increase from the posttest to the transfer test.

To summarize the results, regardless of the difficulty of the topic, the writing scores clearly increased with teacher feedback, both when observed from the overall average and when observed at the individual level. Topic B, which is more difficult, seems to have had a consistent increase in scores due to the lower Pre B task scores at the initial stage, and the drop was smaller than in Topic A, even in the Transfer C task. At the individual level, some students showed differences from the overall group, and the factors that contributed to this may need to be explored.

Figure 4.5 presents four profile plots showing changes over time in total scores on the two tasks for the control and intervention groups. The top two panels show the Pre A, Post A, and Transfer C tasks, and the bottom panels show the Pre B, Post B, and Transfer C tasks.

Figure 4.5



Control Group: Pre-A, Post-A, and Transfer C (n=22) **Task A** *Intervention group: Pre-A, Post-A, Transfer C (n=23)*

Control **Intervention**

Control Group: Pre-B, Post-B, Transfer C **Task B** *Intervention Group: Pre-B, Post-B, Transfer C*

Observing the above four figures, the intervention group shows a slight tendency for data to cluster around a narrow range for the post-intervention measures, with the exception of a couple of students who scored at higher levels. The intervention group appears to have slightly increased their scores compared with the control group, with everyone more linearly aligned. On the other hand, the change over time in the control group appears to be less systematic. Overall, the trajectory of the intervention group line shows a somewhat linear, steadily rising trend. It is also interesting to note that there are a few students in both groups whose scores increased dramatically in the transfer task, and individuals in both groups who started out low ended up showing a V-shaped pattern.

For RQ 2, the scores of the two rewriting tasks and one transfer task were judged to provide meaningful and understandable information for both learners and instructors.

4.1.1.4 RQ 3(1): Examining the Effects of Teacher Feedback Type on Learner Performance on Scores. In response to RQ 3 (1), the effect of the teacher feedback on EFL students' writing in the control and intervention groups, was investigated in terms of the four writing criteria.

4.1.1.4.1 Mixed Between-withing MANOVA. As indicated in Chapter 3, a mixed-between-within MANOVA was conducted to analyze the impact of two different types of teacher feedback on the four criteria in a repeated design. The intervention group (*n* = 23) was given schematized feedback generated by the annotation tool, while the control group (*n* = 22) was given conventional text-based feedback. The four criteria are content, organization, language use, and vocabulary, each scored on a 6-point scale based on the modified ESL CP (Jacobs et al., 1981). A within-group independent variable is occasion: the first draft (pretest), the second draft (posttest), and the transfer task. The between-groups independent variable is the control versus intervention group in terms of teacher feedback. The dependent variables are six levels of scores for the four criteria. Similarly to the previous analyses, Topics A and B, which differ in difficulty, were analyzed separately.

4.1.1.4.2 Preliminary Assumption Checks for Mixed MANOVA. Prior to conducting the MANOVA, preliminary checks were conducted on the preliminary assumptions in terms of the identification of outliers, normality, homogeneity of variance-covariance matrices (based on Box's test of equality of covariance matrices), and the linearity of relationships between the three occasions for each of the four criteria and between the groups (based on regression analysis and scatter plots). Furthermore, the values of bivariate correlations were examined to test the MANOVA assumption that the dependent variables are moderately correlated with each other (Meyer et al., 2006). In other words, a MANOVA should not be performed if there is no correlation between variables or if it includes variables that are not theoretically related (Field, 2009).

Regarding the multivariate outliers for each of the dependent variables for Topic A and Topic B separately, one Mahalanobis outlier was identified in the control group in Topic B, but considering the small sample size and the fact that there were no issues with the Topic A data for the student in question, this student was not excluded from the subsequent analyses. Regarding univariate normality, the Shapiro–Wilk test was conducted. For the control group, some criteria in the Post A tasks were normally distributed ($p > .05$) but others were not ($p < .05$). However, after inspecting the overall shapes of the score distributions in histograms and the Q–Q plots, they appeared to have a distribution that is close to normal. Because a MANOVA is robust to this violation (Pallant, 2011), no steps were taken to address this issue.

Regarding the assumption for homogeneity of variance-covariance matrices, the data for Topic A (Box's M = 136.907, $p = .089$) and Topic B (Box's M = 104.784, $p = .636$) failed to reject the null hypothesis. Therefore, the covariance matrices between the groups were assumed to be equal for the purposes of the MANOVA for Topics A and

B.

Regarding the linearity between the three occasions for each of the four criteria, the scatterplots of the control and intervention groups did not reveal nonlinear trends. Additionally, the scatterplots suggested similar regression slopes in most cases. Therefore, it was assumed that there was no critical violation in terms of linearity.

Lastly, as Table 4.8 shows, the results of bivariate correlations across dependent variables in each task showed reasonable correlations: 92% (41 out of 45) of the correlation coefficients among Topic A and 73% (33 out of 45) of the correlation coefficients among Topic B fell between .2 and .9. These results indicate that a meaningful pattern of correlations was present amongst most of the dependent variables. Hence, MANOVA is appropriate.

Although the normality condition was not necessarily fully satisfied in some cases, given that the conditions were met for the other assumptions, mixed MANOVA was performed for Topics A and B.

Table 4. 8

Pearson Correlations, Means, and Standard Deviations Associated with the Writing Skill Criteria in Two Tasks (n = 45)

Task A

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	*M*	*SD*
A_Con._first	1.000												3.07	0.15
A_Con._second	.635**	1.000											4.09	0.15
A_Con._transfer	0.199	0.257	1.000										4.11	0.16
A_Org._first	.736**	.426**	.442**	1.000									2.96	0.14
A_Org._second	.574**	.694**	0.281	.516**	1.000								4.20	0.15
A_Org._transfer	0.262	0.132	.643**	.364*	0.008	1.000							3.96	0.16
A_Lang._first	.465**	.414**	.304*	.533**	0.265	0.194	1.000						3.16	0.16
A_Lang._second	.354*	.576**	0.284	.357*	.358*	0.270	.662**	1.000					4.22	0.17
A_Lang._transfer	.453**	.380**	.577**	.497**	.439**	.574**	0.230	.473**	1.000				3.36	0.16
A_Voc._first	.417**	.492**	.307*	.468**	.417**	0.104	.733**	.671**	.314*	1.000			3.42	0.15
A_Voc._second	.453**	.690**	.332*	.395**	.576**	0.180	.549**	.662**	.405**	.680**	1.000		4.51	0.14
A_Voc._transfer	.438**	.376*	.546**	.412**	.370*	.519**	0.267	.405**	.766**	.329*	.431**	1	3.40	0.14

Task B

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	*M*	*SD*
1. B_Cont._first	1.000												2.71	0.15
2. B_Cont._second	.586**	1.000											3.91	0.19
3. B_Cont._transfer	0.173	0.216	1.000										4.11	0.16
4. B_Org._first	.668**	.403**	0.127	1.000									2.84	0.17
5. B_Org._second	.479**	.788**	0.077	.533**	1.000								3.93	0.18
6. B_Org._transfer	0.052	0.182	.643**	0.112	0.183	1.000							3.96	0.16
7. B_Lang._first	0.198	0.077	0.099	.326*	0.223	0.175	1.000						2.62	0.15
8. B_Lang._second	0.099	0.263	0.188	.297*	.451**	0.172	.501**	1.000					3.80	0.18
9. B_Lang._transfer	0.259	0.232	.577**	0.198	0.090	.574**	0.264	0.209	1.000				3.36	0.16
10. B_Voc._first	.334*	0.140	0.114	0.280	0.240	0.099	.728**	.402**	.366*	1.000			3.02	0.13
11. B_Voc._second	.305*	0.291	0.243	0.253	.411**	0.244	.428**	.586**	0.234	.569**	1.000		4.22	0.15
12. B_Voc._transfer	0.078	0.184	.546**	0.108	0.047	.519**	0.188	0.176	.766**	0.268	0.122	1	3.40	0.14

Note. ***p*< .01, **p*< .05

4.1.1.4.3 Descriptive Statistics. Table 4.9 shows descriptive statistics of the analytic scores for the four criteria across the three occasions for the two groups given different types of writing teacher feedback. The scores for the more difficult Topic B are lower than those for Topic A. However, the overall trend of the scores for the three occasions is similar for both tasks, which means that the scores increased significantly for the second draft, which is a modified version of the first draft, and decreased to a certain degree for the transfer task.

Table 4.9

Descriptive Statistics in Two Tasks

Task A		Content						Organization					
		first-draft		second-draft		transfer		first-draft		second-draft		transfer	
Group	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Control	22.00	3.05	0.95	4.05	1.13	4.14	1.25	3.00	0.87	4.05	1.21	3.95	0.95
Intervention	23.00	3.09	1.08	4.13	0.92	4.09	0.95	2.91	1.04	4.35	0.83	3.96	1.15

Task A		Language						Vocabulary					
		first-draft		second-draft		transfer		first-draft		second-draft		transfer	
Group	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Control	22.00	2.95	0.90	4.14	1.36	3.32	1.09	3.45	1.06	4.32	1.09	3.14	0.83
Intervention	23.00	3.35	1.15	4.30	0.97	3.39	1.12	3.39	0.99	4.70	0.82	3.65	0.93

Task B		Content						Organization					
		first-draft		second-draft		transfer		first-draft		second-draft		transfer	
Group	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Control	22.00	2.82	1.05	4.00	1.34	4.14	1.25	2.82	0.96	3.73	0.94	3.95	0.95
Intervention	23.00	2.61	0.99	3.83	1.27	4.09	0.95	2.87	1.36	4.13	1.36	3.96	1.15

Task B		Language						Vocabulary					
		first-draft		second-draft		transfer		first-draft		second-draft		transfer	
Group	*n*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Control	22.00	2.64	1.09	3.46	1.14	3.32	1.09	3.05	0.84	4.18	1.01	3.14	0.83
Intervention	23.00	2.61	0.99	4.13	1.22	3.39	1.12	3.00	0.95	4.26	1.05	3.65	0.93

4.1.1.4.4 Mixed MANOVA Results for Analytic Scores. A main analysis of mixed MANOVA was run to investigate the effect of teacher feedback type (conventional versus schematized) on the first draft, second draft, and transfer task scores based on four criteria of the rating scale for Topics A and B. For the main analysis, the scores were compared for the interaction between occasions (repeated measures with three levels) and groups (independent measures with two levels) using MANOVA. Pillai's trace was used to interpret the results because it is supposed to be relatively robust to normality even with small sample sizes (Olson, 1976; Stevens, 1980).

4.1.1.4.5 Topic A Results. Based on Pillai's trace of Topic A, There was no interaction between occasions and groups (F (8, 36) = 1.936, p = .084, η_p^2 = .301), and no main group effect (F (4, 40) = .669, p = .617, η_p^2 = .063), although the main effect of occasions was significant (F (8, 36) = 24.038, p < .001, η_p^2 = .842). Next, the simple main effect of each criterion was examined by univariate ANOVA. Because sphericity was not satisfied based on Mauchly's sphericity test, the Greenhouse–Geisser adjustment for the degrees of freedom was applied for the subsequent univariate tests. As shown in Table 4.10, the occasions factor was significant for all criteria (p < 0.01). As a next step, multiple comparisons were performed for the occasions. For reference, the transition of scores per occasion for each criterion for both groups is shown in Figure 4.6. The analysis yielded four findings: (a) For all four criteria, scores increased significantly from the first draft to the second draft in both the control and intervention groups. (b) There was a significant increase in scores in both groups from the first draft to the transfer task for the content and organization criteria, but not for the language use and vocabulary criteria. (c) For both the control and intervention groups, there was a nonsignificant decrease in the content and organization criteria from the second draft to transfer task, indicating that both skills were retained to some extent. (d) For the language and vocabulary criteria, there was a significant decrease from the second draft to the transfer task and a nonsignificant increase from the first draft to the transfer task.

Table 4.10

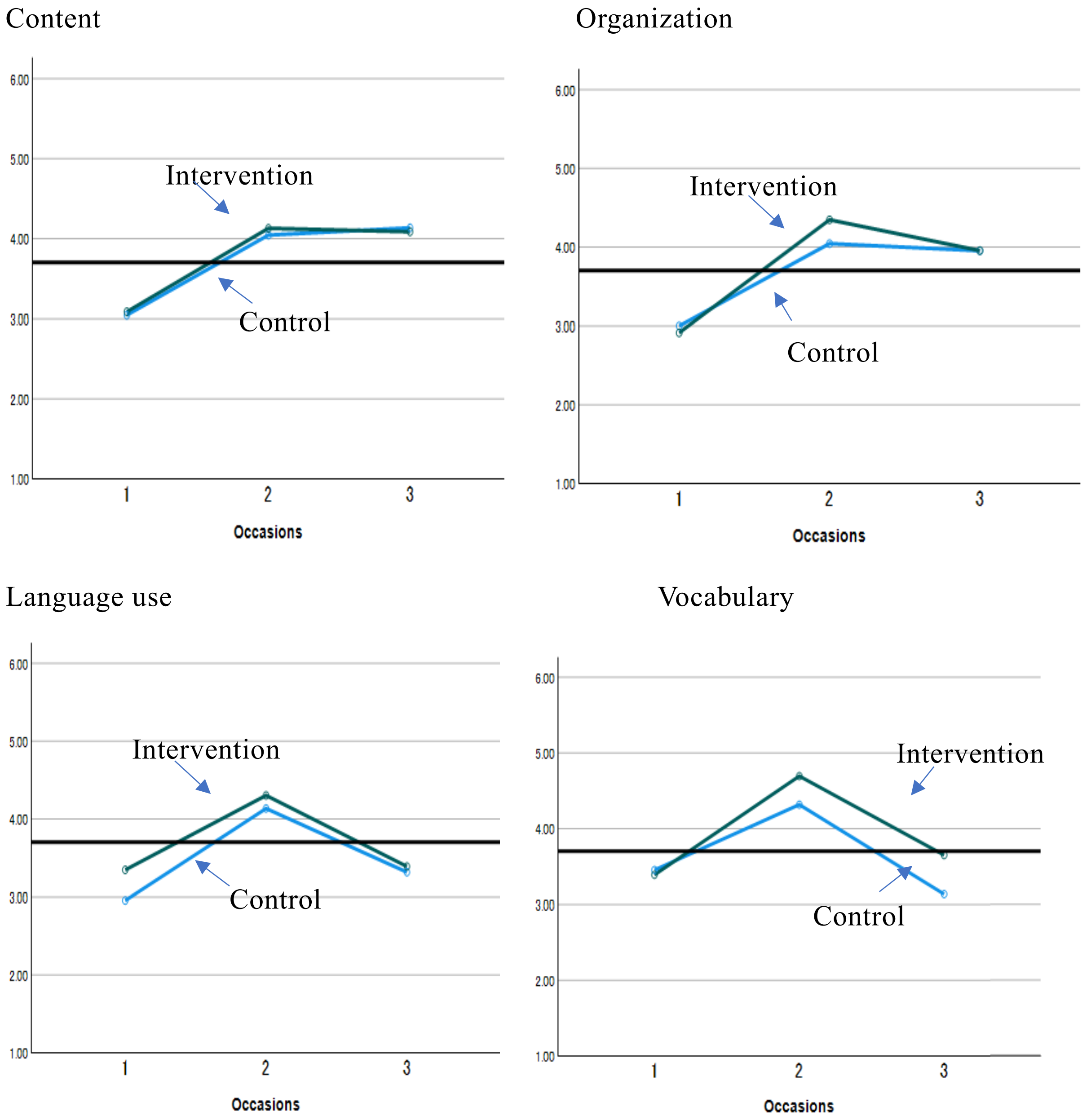
Topic A: Univariate Tests Results

Source	*DV*	*SS*	*df*	*MS*	*F*	*p*	η_p^2
				Between subjects			
Group	Content	.01	1	.01	.01	.91	.00
	Organization	.06	1	.06	.11	.74	.00
	Language	.50	1	.50	.64	.43	.02
	Vocabulary	.86	1	.86	1.42	.24	.03
Residual	Content	27.19	43	.63			
	Organization	23.55	43	.55			
	Language	33.59	43	.78			
	Vocabulary	26.03	43	.61			
				Within subjects			
Occasions	Content	32.05	1.65	19.47	22.52	<.001	.34
	Organization	38.89	1.61	24.20	26.43	<.001	.38
	Language	28.99	1.72	16.86	21.82	<.001	.34
	Vocabulary	36.19	1.75	20.73	38.62	<.001	.47
Occasion × Group	Content	.11	1.65	.06	.07	.90	.00
	Organization	.94	1.61	.58	.64	.50	.02
	Language	.61	1.72	.35	.46	.61	.01
	Vocabulary	2.06	1.75	1.18	2.19	.13	.05
Residuals (occasions)	Content	61.18	70.78	.86			
	Organization	63.27	69.09	.92			
	Language	57.13	73.94	.77			
	Vocabulary	40.30	75.07	.54			

Note. DV, dependent variable

Figure 4.6

Topic A: Comparison of the Estimated Marginal Mean Plots Between the Groups at Three Occasions for Each Criterion



Content

Organization

Language use

Vocabulary

Note. The bold line parallel to the x-axis indicates the overall mean (*M* = 3.7).

4.1.1.4.6 Topic B Results. Based on Pillai's trace of Topic B, there was no main group effect (*F* (4, 40) = .880, *p* = .484, η_p^2 = .081), but there was some interaction between occasions and groups (*F* (8, 36) = 2.399, *p* = .035, η_p^2 = .348). Thus, the simple main effect of the interaction between them was investigated. The language use score for the second draft of the intervention group was significantly higher than that of the control group (p = .03). Furthermore, the intervention group significantly outscored the control group on the transfer task vocabulary (p = .03). Similarly to Topic A, the main effect of occasions was significant (F (8, 36) = 15.944, p < .001, η_p^2 = .780). Subsequently, the simple main effect of each criterion was examined by univariate analysis of variance. Based on Mauchly's sphericity test, sphericity was assumed for the content, organization, and language use criteria, but the Greenhouse–Geisser adjustment for degrees of freedom was applied for the vocabulary criterion in the subsequent univariate tests. As shown in Table 4.11, the occasions factor was significant for all criteria. As a next step, multiple comparisons were performed for the occasions. For reference, the transition of scores per occasion for each criterion and for both groups is shown in Figure 4.7. The analysis yielded three findings: (a) For the control and intervention groups, the scores for all four criteria increased significantly from the first draft to the second draft. (b) In the intervention group, there was a significant increase in the scores of all four criteria from the first draft to the transfer task. However, in the control group, there was a significant increase in scores of the content, organization, and language use criteria, but not the vocabulary criterion, from the first draft to the transfer task. (c) The language scores of the control group changed more moderately, with a smaller increase from the first draft to the second draft than the intervention group, but the decrease in scores from the second draft to the transfer task was also smaller than the intervention group, with no significant difference between the two occasions.

Table 4.11

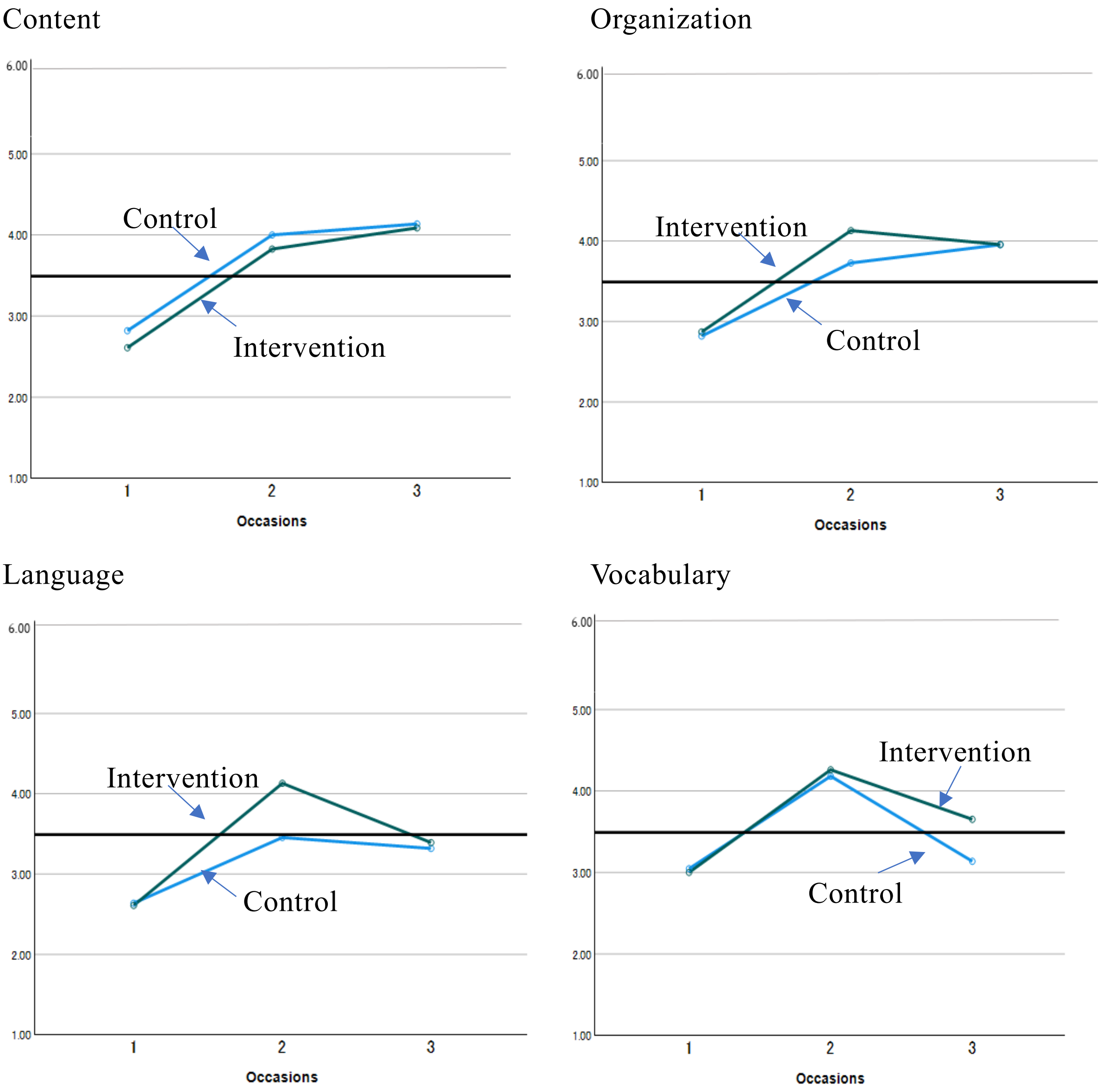
Topic B: Univariate Tests Results

Source	*DV*	*SS*	*df*	*MS*	*F*	*p*	η_p^2
				Between subjects			
Group	Content	.23	1	.23	.32	.57	.01
	Organization	.26	1	.26	.40	.53	.01
	Language	.65	1	.65	.95	.33	.02
	Vocabulary	.38	1	.38	.77	.38	.02
Residual	Content	31.19	43	.73			
	Organization	28.05	43	.65			
	Language	29.33	43	.68			
	Vocabulary	20.99	43	.49			
				Within subjects			
Occasions	Content	51.48	2	25.74	28.78	<.001	.40
	Organization	36.19	2	18.09	20.31	<.001	.32
	Language	31.43	2	15.72	19.06	<.001	.31
	Vocabulary	33.87	1.72	19.64	28.28	<.001	.40
Occasion ×	Content	.16	2	.08	.89	.92	.00
Group	Organization	1.08	2	.54	.60	.55	.01
	Language	3.26	2	1.63	1.97	.15	.04
	Vocabulary	1.95	1.72	1.13	1.63	.21	.04
Residuals (occasions)	Content	76.91	86	.89			
	Organization	76.61	86	.89			
	Language	70.91	86	.83			
	Vocabulary	51.50	74.15	.70			

Note. DV, dependent variable

Figure 4.7

Topic B: Comparison of the Estimated Marginal Mean Plots between the Groups at Three Occasions for Each Criterion



Content

Organization

Language

Vocabulary

Note. A bold line parallel to the x-axis is indicates the overall mean *(M* = 3.5).

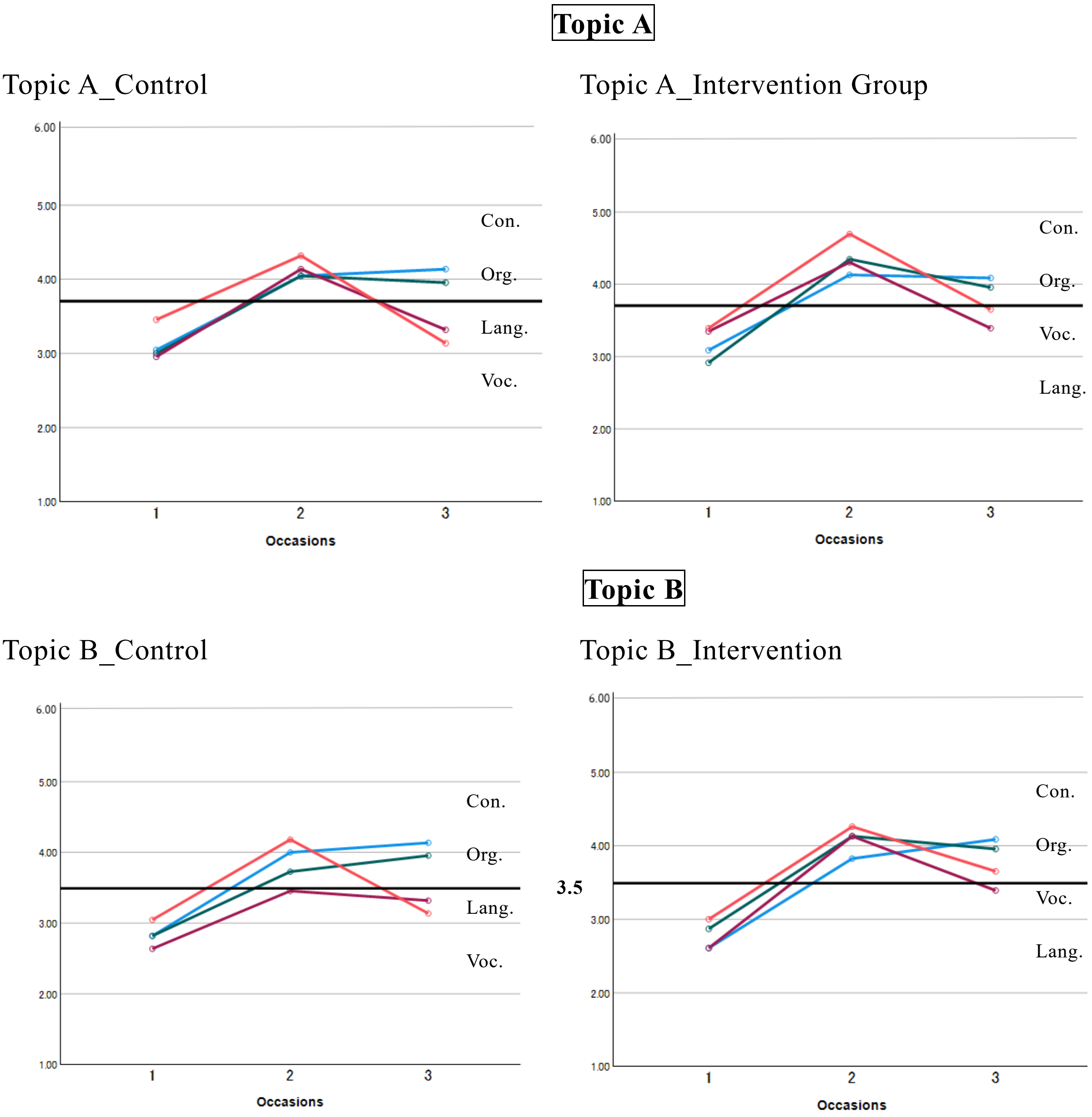
4.1.1.4.7 Control Versus Intervention in Terms of Integration of the Four Dependent Variables. In the previous section, comparisons between groups were made with respect to each criterion. In this section, the four criteria are integrated to provide a more comprehensive perspective to compare the groups. This approach also allows for comparisons between the four dependent variables.

Figure 4.8 integrates the estimated marginal means of the four dependent variables (content, organization, language use, and vocabulary) into a single dimension to present a total of four panels: Topic A_Control and Topic A_Intervention on the upper row, and Topic B_Control and Topic B_Intervention on the lower row. The mean of all dependent variables for each task is also shown as a reference grid (*M* = 3.7 for Topic A; *M* = 3.5 for Topic B).

These diagrams show that (a) in both Topics A and B, based on the reference grid, the intervention group retained higher scores in the transfer ask than the control group. This is more noticeable for Topic B. (b) In the transfer task, regarding the order of scores of four criteria, the order of the lower two was reversed between the groups, with the control group having the order language then vocabulary, whereas the intervention group had the order vocabulary then language. This is true for Topics A and B. (c) Among the four variables, vocabulary seemed to change the most over the three occasions. Namely, scores increased considerably in the second draft, but dropped off markedly in the transfer task.

Figure 4.8

Control Versus Intervention: Estimated Marginal Mean Plots Shift over Three Occasions for the Four Dependent Variables



Topic A

Topic A_Control

Topic A_Intervention Group

Topic B

Topic B_Control

Topic B_Intervention

Note. Con., content; Lang., language use; Org., organization; Voc., vocabulary

4.1.1.4.8 Summary of the Results Obtained from Mixed MANOVA. Based on these results, there were no significant differences for the between-subjects factor, teacher feedback, although the intervention group slightly outperformed the control group in scores throughout the entire study. The effect of the feedback was evident for the three occasions of the within-subjects factor (first draft, second draft, and transfer task). There were significant differences from the first draft to the second draft, with significantly higher scores for the second draft, indicating a significant effect of feedback. Regarding the four criteria, there were several differences in characteristics. In the transfer task, the drop in scores was small for the content and organization criteria, confirming the retention of skills. On the other hand, there was a greater drop in the language use and vocabulary scores, especially in the control group, which seems to be one difference related to the effect of teacher feedback. Moreover, the intervention group tended to retain scores equal to or higher than the overall average on the transfer task compared with the control group.

While there were no significant differences between the intervention group, which was given graphical feedback, and the control group, which was given conventional text-based feedback, in some cases, there were differences in the transfer task or among the criteria. Interestingly, a difference between the two groups in organization was expected, but instead, there was a larger difference in language use and vocabulary. This factor needs to be considered from a comprehensive perspective, including the results of the qualitative study.

4.1.2 Quantitative Analysis Results for Study 2: RQ3(2)

To address RQ 3(2) in Study 2, the relationship between the time required for the process writing revision work and the different forms of teacher feedback was investigated.

4.1.2.1 Comparison of Student Revision Work Time. Table 4.12 presents the relationship between the time required for the process writing revision work and the different forms of teacher feedback. All work was produced using the word-processing software on a computer and submitted to the university web portal, where the time of submission was recorded. The work time was calculated from the submission time.

Table 4.12

The Time on Total Number of Words Written for the Control and Intervention Groups

Group		Control group					Intervention group				
Task		Pre A	Post A	Pre B	Post B	C	Pre A	Post A	Pre B	Post B	C
Mean number of words		122.6	147.5	130.8	148.9	123.3	121.6	144.5	124.9	149.5	129.0
Time (min)	**Mean**	**30.0**	**26.7**	**30.0**	**26.5**	**30.0**	**30.0**	**19.3**	**30.0**	**20.9**	**30.0**
	Median	-	25.5	-	26.0	-	-	16.0	-	22.0	-
	Mode	-	27.0	-	26.0	-	-	16.0	-	17.0	-
	SD	-	6.6	-	7.4	-	-	7.4	-	9.5	-
	Min	-	18.0	-	16.0	-	-	8.0	-	10.0	-
	Max	-	41.0	-	43.0	-	-	35.0	-	44.0	-

Note. There was a 30-minute time limit for the Pre A, Pre B, and Transfer C tasks. The time given for the Post A and B tasks indicates the time the students took to revise with reference to the feedback.

As Table 4.12 shows, the intervention group spent about 7 minutes less time for the Post A task and about 5 minutes less time for the Post B task compared with the control group. For Topic A, which is supposed to be a less difficult task, there was a greater difference between the median and mode values. However, the minimum and maximum values as well as the *SD* are worth noting. In particular, for the Post B task, the intervention group showed a notable difference in the working time. It is not easy to generalize this finding because it may also be related to the students' cautiousness, but given that there was no difference in the number of words produced by the two groups, the intervention group was able to produce the same number of words in a shorter amount of time. Hence, the intervention group was more efficient than the control group.

4.1.3 Summary of the Quantitative Analyses

A series of quantitative analyses were conducted to address RQ 1, a part of RQ 2, and RQ 3 (1) and (2). For RQ 1, an MFRM analysis showed the consistency of ratings (consistency was confirmed as two ratings in pairs, not two raters) in severity. Moreover, the rating scale showed consistent reliability and was thus suitable for the subsequent analyses, although there was some limitation in terms of distance of thresholds in some writing skill criteria. The descriptive statistics of the total scores and parallel coordinate plots showed that the scores on this three-occasion assessment of two rewriting tasks and one transfer task provided meaningful and understandable information for both learners and instructors. Regarding RQ 3 (1), mixed MANOVA revealed no significant difference between the two groups, but in the transfer task, there was a significant increase from the first draft in the intervention group for some of the criteria. Regarding RQ 3(2), the average revision time was shorter for the intervention group. The qualitative study results suggest that schematic feedback may have a greater impact on organization and language, but this eventuality requires more careful discussion by considering the results of the qualitative study discussed in the next section.

4.2 Qualitative Research Results

4.2.1 Qualitative Analyses Results for Study 3-1: RQ3(3)

In Study 3, the students' argumentative writing samples, the same as those analyzed in the quantitative study, as well as the students' open-ended questionnaire responses were analyzed from qualitative perspectives. The qualitative study consisted of one preliminary study followed by two sub-studies (Studies 3-1 and 3-2) that addressed RQ 3 (4), which is related to Claim 1 of Study 3 in the AUA framework (Bachman & Palmer, 2010; Bachman & Damböck, 2017). Part of RQ 3 states: To what extent and how does the type of teacher feedback (conventional versus graphic) affect (3) ideational and rhetorical coherence in the transfer task and (4) students' rewriting behaviors and perception? The first half of the qualitative study in Study 3-1 addresses RQ 3 (3) and Study 3-2 addresses RQ 3 (4).

Before launching the two studies, a preliminary study was conducted to identify and define the rhetorical and ideational coherence in the present study through analysis of anomalies in the students' writing samples to generate the taxonomy by referring to the classification of rhetorical and ideational strategies reported by Barkaoui (2007) and Cumming et al. (2001, 2002). Finally, the findings concerning the two RQs were merged for interpretation with the quantitative results in Studies 1 and 2 to address the overarching hypothesis that the use of schematized feedback is helpful/beneficial to both Japanese EFL students and teachers to teach and learn argumentative stand-alone paragraph writing in terms of ideational and rhetorical coherence.

For the main analysis with the students' writing samples, two researchers, including the author, analyzed 45 transfer tasks by tagging the sentence units using the TIARA annotation tool. The first half of the qualitative study focused specifically on anomalous units that hindered coherence during the tagging process of generating graphical data. Moreover, the product of automatically generated hierarchical tree-view diagrams were examined in terms of their structure and other information provided by the graphics. The second half of the qualitative study involved a thematic analysis of the student responses to the questionnaires on process writing and teacher feedback.

4.2.1.1. A Taxonomy of Ideational and Rhetorical Coherence Summarized for the Study. This section presents a new taxonomy developed by summarizing the definition and rearranged classification of rater judgment strategies based on the identified anomalies and the ESL CP descriptors. Table 4.13 shows a taxonomy of anomalies identified in the actual writing samples categorized into this classification scheme by two annotators. First, the descriptors of the modified ESL CP (Jacobs et al.,1981) used in the quantitative study were classified into corresponding strategies suggested by Barkaoui (2007) and Cumming et al. (2001, 2002). With reference to these classifications of rhetorical and ideational rater judgment strategies in assessing EFL writing, a taxonomy of anomalies was developed by assorting a total of 100 anomalies identified in the 45 students in the Transfer C task writing samples.

Neither Barkaoui (2007) nor Cumming et al. (2001, 2002) defined all the strategies clearly, nor did they distinguish between rhetorical and ideational foci, which are treated as a single category by them, although Cummings et al. (2001, 2002) attached representative comments by the raters in the appendix in their study.[14] Therefore, referring to the descriptors in the ESL CP and the anomalies identified in the samples, nine strategies that were originally considered as a single entity were tentatively divided into four rhetorical and five ideational strategies. In addition, two of the language focus strategies (#10 and #11) were included in Table 4.13 because they were among the anomalies identified in the annotation process. On the other hand, the other seven language focus strategies[15] introduced by Cumming et al. (2001, 2002) were not included in the taxonomy because the problems associated with them are not directly associated with coherence by nature and were not identified as a questionable sentence unit anomaly. The new taxonomy was used in the qualitative analysis of the students' writing samples.

[14] https://www.ets.org/Media/Research/pdf/RM-01-04.pdf

[15] The seven other language focus strategies described by Cumming et al. (2002) that are not included in the analysis of anomalies identified by tagging are as follows: consider gravity of error, consider error frequency, assess fluency, consider lexis, consider syntax or morphology, consider spelling or punctuation, and rate language overall.

Table 4.13

A New Taxonomy of Ideational and Rhetorical Coherence

	Anomalies identified by TIARA annotators	**Corresponding ESL CP descriptors (Jacobs et al., 1981)**	**Raters' judgment strategies (Barkaoui, 2007; Cumming et al., 2001, 2002)**
#	**Elements hindering coherence**	**Organization**	**Rhetorical focus**
1	・Improper order of ideas ・Absence of connecting ideas ・Leap of logic ・Multiple ideas in a sentence ・Convoluted ideas	・Are the points logically developed, using a particular sequence such as time order, space order, or importance? ・Is this development indicated by appropriate transitional markers? ・Are there effective transition elements—words, phrases, or sentences—that link and move ideas within the paragraph? ・Do the ideas flow, building on one another? ・Is the overall relationship between sentences clearly indicated? ・Is enough written to adequately develop the subject?	Assess reasoning, logic, or topic development
2	・Deviation from the topic; the presence of an out-of-step (or off-topic) sentence unit	・Does each group of sentences reflect a single purpose? ・Is there a clearly stated controlling idea or central focus to the paragraph? ・Do sentences support, limit, and direct the topic sentence? ・Do the group of sentences form a unified writing? ・Are all ideas directed concisely to the central focus of the paragraph, without digressions?	Assess coherence
3	・Redundant ideas	(None)	Identify redundancies
4	・Lack of organizational constituent(s)	・Are there a topic sentence and concluding sentences? ・Is there a beginning, a middle, and an end to the writing?	Assess sentence organization
	Anomalies hindering coherence	**Content or vocabulary**	**Ideational focus**
5	・Misunderstanding of the task	・Is there understanding of the subject? ・Is the thesis expanded enough to convey a sense of completeness?	Assess task completion
6	・Inappropriate support due to insufficient knowledge on the subject	・Are facts or other pertinent information used? ・Is all information clearly pertinent of the topic? ・Is extraneous material excluded?	Assess relevance

7	N/A	・Is there originality with concrete detail to illustrate, define, compare, or contrast factual information supporting the thesis?	Assess interest, originality, or creativity
8	・Insufficient understanding of rebuttal	・Is there recognition of several aspects of the subject? ・Are the interrelationships of these aspects shown? ・Are several main points discussed? ・Is there sufficient detail? ・Is there a specific method of development (such as comparison/contrast, illustration, definition, example, description, fact, or personal experience)? ・Is there an awareness of different points of view?	Rate ideas or rhetoric
9	N/A	Corresponding descriptor is included in the vocabulary category	Assess style, register, or genre
	Anomalies identified by annotators	**Vocabulary or language use**	**Language focus**
10	・Insufficient sentence length as a paragraph	・Not enough to evaluate	Assess quantity of total written production
11	・Stagnation due to an incomprehensible sentence	・Does not communicate (for language use) ・Little knowledge of English vocabulary (for vocabulary)	Assess comprehensibility

Each category in Table 4.13 can be explained and defined by using some of the rater comments presented by Cumming et al. (2002) in the appendix as a reference for some definitions. Furthermore, to define the concept, descriptors in EFL essay writing proposed by Matsumura and Takagi (2022) were used.

Categories 1–4 were grouped under the scope of rhetorical focus. The anomalous elements in Category 1 are based on the two raters' judgment in the present study to evaluate the sentence logic flow created by proper sequence and transition elements, which is assumed to correspond to judgment strategy of "assess reasoning, logic, or topic development of raters" in the study by Cumming et al. (2001, 2002) as well as the descriptors of the ESL CP (Jacobs et al., 1981) shown in the adjacent cells in Table 4.13. For example, improper order of ideas, absence of connecting ideas, leap of logic, multiple ideas in a sentence, and convoluted ideas belong to Category 1. The elements in Category 2, presumably corresponding to "assess coherence of raters" in the judgment strategy, are based on the decision-making to assess propositional consistency throughout the paragraph based on the interpretation of other sentences. The element identified as deviation from the topic and the presence of an out-of-step (or off-topic) sentence unit fall into Category 2. Among Categories 1–4, the first two are collectively defined as centrally related to van Dijk's (1977) interpretation of coherence as "semantic property of discourses, based on the interpretation of other sentences" (p. 93). Category 3, which corresponds to identify redundancies, is based on the raters' evaluating behavior to locate unnecessary repetition in the sentence. Category 4 is based on the raters' decision regarding the presence or absence of organizational components in students' writing, such as topic, supporting, and concluding sentences, which is regarded to correspond to "assess sentence organization" in the study by Cumming et al. (2001, 2002).

Categories 5–9 represented the ideational focus in the present study. Category 5 is based on the assessment to determine a writer's proper understanding of the purpose or intent of the assignment, or whether the assignment is properly completed. This category corresponds to assess task completion of the raters' judgment strategy. Category 6, assess relevance, is based on the judgment whether irrelevant and off-target statements are eliminated. Category 7, assess interest, originality, or creativity, evaluates "uniqueness; intriguingness; empathy with readers" of writing products (Matsumura & Takagi, 2022, p. 51). However, there was no applicable element in this study's classification of anomalies focusing on coherence, although it is listed as a descriptor in the ESL CP or as one of the rhetorical and ideational strategies and is related to rater's evaluation of the content scores. This decision was made because this category has little to do with the rater's determination of the presence or absence of anomalies. This also applies to Category 9, assess style, register, or genre, which can be defined as a "bookish version of English; [a] formulaic way of writing composition; appropriateness to the situation or type of writing" (Cumming et al., 2001, p. 9, 2002, p. 94). Therefore, Categories 7 and 9 were not subject to the anomaly analysis. Category 8 requires a more careful description. In the present study, the related anomalous elements were identified by focusing on a specific combination or relationship between sentence units in terms of argument. In other words, the judgment standard of the category is defined as balanced and elaborate rhetoric carried throughout the paragraph including a successful/well-embedded rebuttal. Briefly, it is a question of the success or failure of the rebuttal. The category is assumed to corresponds in part to rate ideas or rhetoric, which is described as being "not rhetorically very sophisticated" (Cumming et al., 2001, p. 93, 2002, p. 94). The judgment for the elements in the category wants "comprehensive judgement" when "it is difficult to make a partial decision" (Matsumura & Takagi, 2022, p. 51).

Lastly, Categories 10 and 11 represent the language focus; they try to locate the language-related problems. They are often found especially in products by learners with low English proficiency and their impact on the rating of writing is not negligible. For Category 10, although none of the writing samples in this study was applicable, which is defined as being inaccessible due to a significantly insufficient word count, it was left as a possible standard for inappropriateness of condition for coherence of passage. Category 11 is well represented by the descriptors of the ESL CP shown in Table 4.13.

Specific examples from the student writing samples for each anomaly category are presented in Section 4.5.1.3.

4.2.1.2 Analysis of Anomalies Identified by the Annotators with the Annotation Tool. This section presents the qualitative analysis of argumentative stand-alone paragraphs produced by 45 students (22 from the control group and 23 from the intervention group) at the final phase of process writing in the classroom, tagged and schematized with the TIARA annotation tool for the Transfer C task. The transfer writing task was assigned after the students had completed two paragraphs on two topics, which differed in difficulty, and revised the paragraphs based on the teacher feedback, to investigate the retention of the effects of previous learning.

The analysis focused on sentence units that the annotators identified as anomalies because anomalous units have a significant bearing on the coherence of a passage in that they reduce the flow of the entire paragraph. This analysis allowed examining the differences in the above aspects between the control and intervention groups, and to investigate the effects of writing feedback with diagrams, which are difficult to recognize from the scores alone. The effect of feedback in process writing requires quantitative analysis of scoring results as well as detailed sentence analysis. Especially in classroom assessments, such analysis is required to identify qualitative changes in student writing products, which may not necessarily be recognizable solely with quantitative analysis. Classroom teachers' interest, particularly in formative assessment, is not directed at rank ordering, but rather at student growth and goal achievement.

4.2.1.3 Relationship Between Organizational Elements and Anomalies for All Student Transfer Task Writing Samples. Tagging sentences with TIARA allows labeling organizational elements and locating problematic units as organizational and/or ideational anomalies. This endeavor allows exploring with which parts of opinion paragraphs, including the counterargument and rebuttal approach, students tend to have these problems. Analyzing the content of these problems and how they can be categorized can reveal some of the characteristics of paragraphs written by Japanese EFL students.

The exact agreement between the two annotators on the newly developed taxonomy (Table 4.13) was 0.71. The degree of agreement for decisions on the type of anomalies was lower than that for decisions on location because multiple interpretations were possible. In cases where there were discrepancies in the other labels, the two annotators discussed them together and finally reached an agreement. The tally in terms of anomaly frequencies and the total units in each organizational elements are summarized in Table 4.14. The annotators identified 100 anomalies in the 45 writing samples. The percentage of anomalies in the 453 total sentence units was 22.1%. Among them, rebuttal (39.3%) had the highest percentage of anomalies, followed by detail (26.2%), support (23.9%), concluding sentence (8.9%), and topic sentence (6.7%).

Table 4.14

The Frequencies of Anomalies in the Organizational Element in the Transfer Task

Elements in argument	Error frequencies	Total units	Proportion of error units per element
Topic sentence	3 (1:2)	45	6.7%
Support	27 (13:14)	112	24.1%
Detail	49 (31:18)	194	25.3%
Counterargument	4 (4:0)	27	14.8%
Rebuttal	13 (7:6)	30	43.3%
Concluding sentence	4 (3:1)	45	8.9%
Total	100	453	22.1%

Note. The analysis is based on 45 writing samples. The ratios in parentheses indicate the frequency in the control versus intervention group. For example, the total number of 45 concluding sentences comprises 41 instances of concluding sentences without problems and four instances of non-existent elements that should be present.

A detailed analysis is provided later in this section, so a brief description of Table 4.20 is provided here. Although there were fewer problems in topic sentence and concluding sentence than in the other elements, the significance of the problem is not necessarily small, especially with the cases in the topic sentence category. The impact of a misstep at the starting point that sets the direction of the entire sentence is not small. This could result in low organization scores. Detail had the largest number of problematic units because the total number of details themselves is quite large. In terms of the ratio of anomalies to the total number of elements, it is about the same as support, both being around 25%. For rebuttal, 13 out of 30 units (almost 40%) were problematic, indicating that the students were not successful in incorporating a rebuttal into their paragraphs. This difficulty was also found in the pilot empirical study with high school students by the author (Matsumura & Sakamoto, 2021 p. 40). This trend may be common in opinion paragraph writing by Japanese learners of English, at least at the beginner-intermediate level.

4.2.1.4 Probable Cause, Location, and Frequency of Anomalous Units Tagged as "Questionable" (Control Versus Intervention). Sentences tagged as "Questionable," representing a type of anomaly, were examined in terms of the elements of the paragraph in which they occurred. Table 4.15 shows the student writing samples in the rows and the frequencies and types of the probable causes in each element in the columns. The italicized numbers in the "cause(s)" column correspond to the 11 categories shown in Table 4.13. Sample numbers 101–124 represent the control group, and 201–227 represent the intervention group. The numbers in the columns of the six elements indicate how many anomalies were found within each element. For example, the writing sample for Student 103 has a Category 1 anomaly (a problem with reasoning, logic, and topic development) in one of the supports, two Category 1 anomalies in the three detail units, and one Category 3 anomaly (deviation from the topic). The shaded student number cells—for example, 107, 109, and 114— indicate that the two annotators did not identify organization problems.

Table 4.15

Frequency Matrix of Probable Causes for Anomalous Unit in Organization

Ss		tp	*cause(s)*	sup	*cause(s)*	det	*cause(s)*	ca	*cause(s)*	reb	*cause(s)*	con	*cause(s)*
Control Group	101							1	*1*				
	102									1	*8*		
	103			1	*1*	3	*1,1,3*						
	104					1	*11*						
	105					1	*6*	1	*1*	1	*8*		
	107	Shaded Ss cells were no anomaly identified.											
	108			1	*1*	7	*1,1,1,1,1,1,1*						
	109												
	110			1	*1*	1	*6*	,		1	*8*		
	111			1	*3*	1	*11*	1	*1*	1	*8*	1	*4*
	112	1	*4*			1	*1*	1	*6*				
	113			1	*1*	1	*6*						
	114												
	115												
	116			2	*1,1*	5	*1,1,6,6,6*						
	117			2	*1,11*	1	*11*						
	118					1	*6*			1	*8*		
	120			1	*6*					1	*8*		
	121			1	*1*	1	*6*					1	*4*
	122					1	*1*						
	123			1	*11*	5	*6,6,6,6,11*			1	*8*		
	124			1	*6*	1	*11*					1	*4*
Intervention Group	201	1	*3*	1	*6*	2	*3,11*			1	*8*		
	203					1	*1*						
	204												
	205			1	*11*	1	*1*					1	*4*
	206			1	*6*	1	*1*						
	207			1	*11*								
	208			1	*1*	1	*2*			1	*8*		
	209			1	*6*								
	210			1	*6*								
	211									1	*8*		
	212												
	213			1	*11*								
	214			1	*11*	4	*1,3,11,11*						
	215												
	216												
	217					2	*1,4*						
	218					1	*1*			1	*8*		
	219			1	*6*					1	*8*		
	220	1	*4*	1	*11*	2	*1,6*						
	221												
	225			1	*1*					1	*8*		
	226												
	227			2	*2,3*	3	*1,1,6*						

Note. The table provides a comparison between the control and intervention groups. The numbers in italics in the “cause(s)” cells correspond the 11 anomaly categories in Table 4.13. The number in the cell of the element indicates the frequency of the anomaly. Ss, students; tp, topic sentence; sup, supporting sentence; det, detail; ca, counterargument; reb, rebuttal; and con, concluding sentence

Comparing the control and intervention groups yielded several findings. First, the control group had only four “anomaly-free” writing samples (18% of the group), while the intervention group had seven (30% of the group). Hence, in the transfer task more students in the intervention group seemed to be successful in producing anomaly-free paragraphs in terms of the rhetorical and ideational aspects. Second, there were four instances of anomalies in counterargument in the control group and zero in the intervention group. This difference is noteworthy when compared with the almost equal number of six and five instances for both groups in rebuttal. This finding indicates that it was equally difficult for both groups to incorporate rebuttal elements successfully into a paragraph, but at the same time, the intervention group showed a retention effect based on successfully integrating a counterargument element into the sentence. This phenomenon may be attributed to the fact that counterarguments need only present a viewpoint that differs from their own, but rebuttals are expected to be more difficult because the writer is required to further refute it and to be consistent. It is also notable that detail anomalies occurred more frequently in the control group than in the intervention group, and that some writing samples had a series of anomalies within a single element. This finding may also suggest that the students in the intervention group were slightly more attentive to consistency in paragraph detail than the students in the control group. Moreover, only one sample in the intervention group lacked a concluding sentence, while three in the control group had an issue with the concluding sentence. Again, this can be seen as a sign of some awareness of the students in the intervention group regarding the importance of remembering to include the organizational components. Lastly, the number of anomalies in the support position did not differ between the two groups. However, when examining the causes, the intervention group showed either Category 6 (problem with relevance) or Category 11 (comprehensibility) errors, whereas the control group had more rhetorical errors, specifically in Category 1 (a problem with reasoning, logic, topic development). Hence, the intervention group seemed to include fewer rhetorical anomalies than the control group in the transfer task.

Interpreting these results requires caution because the small overall sample size make it difficult to generalize the results, although the inferences described above may be drawn.

4.2.1.5 Causes of Anomalies (Control Versus Intervention). Table 4.16 shows the breakdown of frequencies of anomalous sentence units in the control and intervention groups divided by the 11 anomaly categories that hinder coherence.

Table 4.16

Breakdown of Anomaly Frequencies by the 11 Anomaly Categories by Group

#	Anomaly category	Frequency of anomalous unit identified		
		Control	Intervention	Total
1	Poor reasoning, logic, topic development	24	10	**34**
2	Deviation from the topic	0	2	**2**
3	Redundancy	2	4	**6**
4	Lack of organizational element	4	3	**7**
	Subtotal	**30**	**19**	**49**
5	Task incompletion	0	1	**1**
6	Irrelevant support	15	7	**22**
7	Originality	-	-	**-**
8	Insufficient rebuttal	7	6	**13**
9	Style, register	-	-	**-**
10	Quantity	0	0	**0**
11	Incomprehensibility	7	8	**15**
	Subtotal	**29**	**22**	**51**
	Total	**59**	**41**	**100**

Note. Categories 1–4 are the rhetorical focus, Categories 5–9 are the ideational focus, and Categories 10 and 11 are the language focus. The boxes in the square indicate data that are specifically mentioned in the main sentence.

There was a difference between the control and the intervention groups in the total frequency of anomalies, 59 instances for the control group and 41 instances for the intervention group. The chi-square goodness-of-fit test was used to determine whether the difference between the groups was significant (selected because the categorical variable is 1). The tested yielded $\chi^2(1, N = 100) = 3.24$, $p = .07$, meaning the difference between the groups was not significant (although there is a trend for a difference). The relatively low number of anomalies in the intervention group can be attributed in part to the fact the larger number of anomaly-free writing samples compared with the control group. In fact, it should be noted that the ratio of anomalous units per sample, excluding anomaly-free samples, was 3.3 instances in the control group and 2.7 instances in the intervention group, a 20% reduction in the intervention group. This finding suggests that even in the intervention group, there may be a division between students who exhibit the retention effect and those who do not.

Next, a chi-square test was conducted to examine whether there was a significant difference in the frequency of occurrence between the control group and the intervention group in the estimated causes of anomaly types, excluding three items that were zero in both groups. The result was $\chi^2(7, N = 100) = 9.70$, $p = .19$, Cramer's $V = .31$, meaning there was no significant difference between the groups. It should be noted that the control group had twice as many Category 1 and Category 6 anomalies as the intervention group.

Because Category 1, which accounts for 34% of the total frequency, includes several types of anomalies, further detailed examination was necessary. Table 4.17 shows a breakdown of anomalous unit frequencies for Category 1 sorted by the five sub-categorized anomalies, comparing the control and the intervention groups. The most striking between-group difference was in the frequency of improper order of ideas, while the rest showed almost no difference between the groups. Although the difference between the groups was not significant, the large numerical difference between the groups (13 for the control group and 2 for the intervention group) cannot be ignored.

Table 4.17

Further Breakdown of Anomalous Unit Frequencies for Category 1

Corresponding judgment strategy	Anomalies identified by the annotators	Control	Intervention
1. Assess reasoning, logic, or topic development	Improper order of ideas	13	2
	Absence of connecting ideas	2	2
	Leap of logic	1	0
	Multiple ideas in a sentence	2	1
	Convoluted ideas	6	5
	Total	24	10

Note. The box in the square indicates data that are specifically mentioned in the main sentence.

4.2.2 Linking the Quantitative and Qualitative Results for Interpretation

4.2.2.1 Relationship Between Anomalies and Analytic Writing Scores Focusing on the Organization and Content Criteria. The relationship between the frequency of anomalous units identified as "questionable" by the annotator in the TIARA tagging process and the scores on the modified ESL CP analytic rating scale was explored. Between the content, organization, language use, vocabulary, and mechanics criteria, the focus was on the organization and content criteria, which should be directly related to TIARA tagging, whose analysis is based on rhetorical and ideational aspects of writing.

4.2.2.1.1 Relationship Between the Organization and Content Criteria in Terms of the Frequency of Anomalies. The earlier quantitative analysis with an MFRM analysis confirmed the unidimensionality of the data. It is necessary to examine the relationship between the organization and content criteria specifically in terms of frequency of anomalies. Table 4.18 shows the correlation between the total frequency of anomalies in the writing samples and the content and organization scores. The two criteria are correlated at a Pearson correlation coefficient of .70, but the small sample size (*n* = 45) means that this statistic needs to be interpreted carefully. Still, it should be fine to conclude that they are correlated to some degree. Each of the two rating scale criteria and the frequency of anomalies were moderately negatively correlated, with a Pearson correlation coefficient of -.497 and -.511, respectively. It should also be noted that the two values are quite similar to each other. They show inverse relationships, where both organization and content scores tended to decrease as the frequency of anomalies increased. These values are considered large enough because the number of anomalies in one paragraph is not the only factor involved in the score; the presence of other factors can also be inferred. Based on these results, it was assumed that the organization and content scores and the frequency of anomalies could be discussed on the same dimension. Therefore, there would be no problem to proceed with the analysis using a score ranking that integrates the organization and content scores aligned with the qualitative perspective of frequency of anomalies identified on rhetorical and ideational perspectives.

Table 4.18

Correlation Between the Total Frequencies of Anomalies and the Content and Organization Scores

	Content score	Organization score	Total anomalies
Content score	1.00		
Organization score	.697**	1.00	
Total anomalies	-.497**	-.511**	1.00

***p* <.01, *df* = 43

4.2.2.2 Relationship Between the Location of Occurrence of Organizational Elements of Anomalies and the Organization and Content Scores. As shown in Section 4.2.2.1.1, there was a negative correlation between the frequency of anomalies and the organization and content scores. The relationship between the scores and the anomaly categories in students' writing was explored to determine whether other factors besides the number of anomalies could affect the scores. In other words, the point of interest was which organizational elements identify anomalies.

Table 4.19 shows the 45 writing samples sorted by the analytic score for organization and then by content (25 points in total) in descending order. The reason for sorting by the 25-point scores rather than by the 6-point scores is to provide a more detailed rank order. Additionally, the 6-point scores of this study and the four-level ordinal scale classification following the original rating scale of Jacobs et al. (1981) are also included. The frequency of occurrence of anomalies for each sample is listed, and the anomaly-free cells in the total anomalies' column are shaded. In addition, cells where anomalies occurred are colored for visibility. The total number of words and units (i.e., sentences) of individual paragraphs are included as additional information. This addition was considered necessary because as the number of words increases, the possibility of the risk of error/anomaly occurring may, of course, also increase, and vice versa. The number in the cell under the name of frequencies of anomalies is the number of anomalous units identified for each element.

Table 4.19

Relationship Between the Anomaly Categories and Organization (Considering Content) Scores in Descending Order and the Evaluation Level

Ss	wds	total units	Cont. score	Org. score	Total anomalies	Frequenceis of anomalies						6-level scale	ESL 4-level ordinal scale
						tp	sup	det	ca	reb	con		
216	167	10	22	24	0							**6 (over 22)**	**Very good**
122	136	10	20	23	1			1					
209	127	10	20	22	1		1						
120	143	9	22	21	2		1			1		**5 (20-21)**	**Good to average**
121	114	10	22	21	3		1	1			1		
113	123	11	21	21	2		1	1					
117	100	8	21	21	3		2	1					
211	118	9	21	21	1					1			
212	120	10	20	21	0								
103	106	11	19	21	4		1	3					
210	139	11	19	21	1		1						
203	141	11	20	20	1			1					
218	118	9	20	20	2			1		1			
206	102	9	18	20	2		1	1					
114	157	10	22	19	0							**4 (18-19)**	
102	136	10	20	19	1					1			
115	122	9	19	19	0								
213	106	8	19	19	1		1						
105	152	9	18	19	3			1	1	1			
107	88	8	17	19	0								
109	125	8	20	18	0								
204	127	13	20	18	0								
118	100	9	18	18	2			1		1			
219	127	8	18	18	2		1			1			
225	128	11	18	18	2		1			1			
226	132	10	18	18	0								
111	130	8	17	18	5		1	1	1	1	1		
108	114	11	16	18	8		1	7					
215	109	9	16	18	0								
110	151	11	20	17	3		1	1		1		**3 (15-17)**	**Fair to Poor**
208	155	11	19	17	3		1	1		1			
104	141	12	18	17	1			1					
207	128	9	16	17	1		1						
221	98	9	16	16	0								
112	113	7	12	16	3	1		1	1				
205	101	6	20	15	3		1	1			1		
101	106	8	18	15	1				1				
124	129	10	18	15	3		1	1			1		
217	166	10	17	15	2			2					
227	124	10	15	15	5		2	3					
116	122	11	14	15	7		2	5					
201	146	13	18	13	5	1	1	2		1		**2 (11-14)**	
220	166	13	12	12	4	1	1	2					
214	123	12	18	11	5		1	4					
123	104	14	11	11	7		1	5		1			
No applicable writing samples at this level												**1 (under 10)**	**Very poor**

Note. Ss, students; tp, topic sentence; sup, supporting sentence; det, detail; ca, counterargument; reb, rebuttal; and con, concluding sentence.

Moving on to observations on each element, regarding anomalies in the topic sentence, all three instances occurred at score levels 3 and 2, or fair to poor. In contrast, there were anomalies in the concluding element at score levels 5, 4, and 3, relatively higher score levels than those involving the topic sentence anomalies. This may be due to the fact that errors in the topic sentence are more critical and lead to lower scores, whereas problems such as missing a concluding sentence do not necessarily lead to comprehension problems. There were no counterargument anomalies at the score levels 5 and 6; they appeared for the first time in the middle of level 4. In other words, the high-scoring samples had no counterargument anomalies. Hence, samples that present problems in the counterargument were less likely to be highly rated. In contrast, there were rebuttal anomalies at all score levels. This result suggests that although the counterargument and the rebuttal should appear in pairs, it is presumably more difficult to succeed in incorporating a rebuttal effectively that further adapts to the existing trend against a counterargument, which only needs to present a view that differs from the writer's own assertion. In the open-ended questionnaire, several students remarked about the difficulty in incorporating a rebuttal into a paragraph (discussed later in this chapter). Lastly, one thing to note here is about the anomaly-free writing samples. In many cases, these samples scored higher, but in some cases, such as Students 107, 215, and 221, there were fewer words in this paragraph. This factor may have avoided errors/problems identified. Note that in those cases, the lack of anomalies did not necessarily result in a higher score from the overall assessment.

4.2.2.3 Relationship Between the Causes of Anomalies and the Organization and Content Scores. Next, the relationships between the scores and the 11 categories of causes classified by the annotators were evaluated. Table 4.20 shows the data arranged in the same order as Table 4.19, with the anomaly causes in the writing sample listed in descending order by the organization and content scores. The results and interpretations are presented one by one for Categories 1–11. However, note that while Categories 7–9 were included in the scoring scale, they were not included in the analysis based on anomalies, and Category 10 was not applicable to the actual writing samples in the present study. Therefore, the columns for these three items are blank. The explanation of each anomaly element includes an excerpt from a writing sample for illustration.

Table 4.20

Relationship Between Anomaly Types and Organization (Considering Content) Scores in Descending Order and Evaluation Level

Ss	wds	total units	Cont. score	Org. score	Frequencies of anomalies by cause											6-level scale	ESL 4-level ordinal scale
					#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11		
216	167	10	22	24							-		-			**6 (over 22)**	**Very good**
122	136	10	20	23	1						-		-				
209	127	10	20	22						1	-		-				
120	143	9	22	21						1	-	1	-			**5 (20-21)**	**Good to average**
121	114	10	22	21	1			1		1	-		-				
113	123	11	21	21	1					1	-		-				
117	100	8	21	21	1						-		-		2		
211	118	9	21	21							-	1	-				
212	120	10	20	21							-		-				
103	106	11	19	21	3		1				-		-				
210	139	11	19	21						1	-		-				
203	141	11	20	20	1						-		-				
218	118	9	20	20	1						-	1	-				
206	102	9	18	20	1					1	-		-				
114	157	10	22	19							-		-			**4 (18-19)**	
102	136	10	20	19							-	1	-				
115	122	9	19	19							-		-				
213	106	8	19	19							-		-		1		
105	152	9	18	19	1					1	-	1	-				
107	88	8	17	19							-		-				
109	125	8	20	18							-		-				
204	127	13	20	18							-		-				
118	100	9	18	18						1	-	1	-				
219	127	8	18	18						1	-	1	-				
225	128	11	18	18	1						-	1	-				
226	132	10	18	18							-		-				
111	130	8	17	18	1		1	1			-	1	-		1		
108	114	11	16	18	8						-		-				
215	109	9	16	18							-		-				
110	151	11	20	17	1					1	-	1	-			**3 (15-17)**	**Fair to Poor**
208	155	11	19	17	1	1					-	1	-				
104	141	12	18	17							-		-		1		
207	128	9	16	17							-		-		1		
221	98	9	16	16							-		-				
112	113	7	12	16	1			1		1	-		-				
205	101	6	20	15			1	1			-		-		1		
101	106	8	18	15	1						-		-				
124	129	10	18	15				1		1	-		-		1		
217	166	10	17	15	1			1			-		-				
227	124	10	15	15	2	1	1			1	-		-				
116	122	11	14	15	4					3	-		-				
201	146	13	18	13			1			1	-	1	-		1	**2 (11-14)**	
220	166	13	12	12	1			1	1	1	-		-		1		
214	123	12	18	11	1		1				-		-		3		
123	104	14	11	11						4	-	1	-		2		
No applicable writing samples at this level																**1 (under 10)**	**Very poor**

Note. The 11 anomaly categories are defined in Table 4.13

4.2.2.3.1 Category 1: Reasoning, Logic, or Topic Development. Category 1 anomalies occurred evenly at all scoring levels. One of the anomalous elements, improper order of ideas, tended to identify multiple anomalies within a single writing sample, as can be seen in those of Students 103, 108, and 116 (Table 4.20), because this problem involves multiple adjacent sentence units, such as sentence ordering issues in one paragraph. It should be noted that the three aforementioned writing samples are from the control group.

To be more specific, an example from Student 108 is presented below followed by an explanation. The anomalies discussed are underlined. Note the following abbreviations in the example sentence, which also apply to the subsequent examples: TP, topic sentence; SUP, support; DET, detail; CA, counterargument; REB, rebuttal; CON, concluding sentence; ?, questionable sentence/anomaly.

> **Example 1 (#1-1): Improper order of ideas** (Student 108)
> TP[1] I think it is beneficial for workers to change jobs often. ←DET [2] I have a reason for this idea. ←? [3] The life that we can experience is only one. ←? [4] Work is the biggest part of the life. ←? [5] In Japan, we are called workers from 22 to 65 years old. ←? [6] We have to be workers about half of the life. ←? [7] Also, we have less body limits to do what we want before you become to be old. ←? [8] Do you want to use the biggest part of the life for what you don't want to do? ←? [9] The answer is "No." ←? [10] The best choice is that doing anything you want. ←CON [11] For this reason, I think changing job is beneficial for workers.

Example 1 includes sentences 3–10. The writer is in favor of workers changing job often and she declares in sentence 2 that she will now give her reason. However, the next sentence, "The life that we can experience is only one," puzzles the reader. After reading the subsequent sentences, the reader gradually understands how such a background description could support the writer's claim. Starting with a background explanation often confuses the reader. This kind of anomaly can be addressed by connecting some sentences together or by reordering, for example, by showing the landing point at the beginning.

Example 2 from Student 203 represents absence of connecting ideas.

Example 2 (#1-1): Absence of connecting ideas (Student 203)
←SUP [3] First, worker should find jobs that they want to have. ←DET [4] If workers hate their jobs, jobs will give them negative effects such as sick. ←DET [3] However, if they have jobs that they want. It will give them a lot of positive effects. ←? [6] They should find and change their jobs that they want to have.

In this example, the writer explains step by step, with conditions, that if you are in a job you do not like, you should change jobs. In this case, it is important to clarify this "condition" in the argument, but it is not clearly stated in sentence 6. The anomaly in this case is not so serious that it reduced the score. However, it would be more effective to add a limitation in sentence 6 more clearly as a connecting idea to the preceding sentences, such as "if the workplace is not a good fit" or "in order to avoid negative personal influences."

Example 3 shows a leap of logic. Student 101 believes that it is not beneficial for workers to change jobs often.

Example 3 (#1-3): Leap of logic (Student 101)
TP [1] I think it is not beneficial for workers to change jobs often. ←SUP [2] Because to change jobs can lose a steady income. ←DET [3] If that happens, our lives will be in danger. ←? [4] On the other hand, some would say it is a way out of hard work situation. ←SUP [5] And to change jobs can lead to improvement the present work situation. ←REB [6] But that is not always guaranteed.

There are two problems with sentence 4. One is that the pronoun "it" has jumped over the immediately preceding "income"-related content that it should have succeeded due to a leap in content, making it ambiguous, though a reader familiar with the EFL error may probably assume readily that it refers to a "job change." Another problem is "On the other hand" used as a connective, which is not directly in contrast to income, so the use of this phrase would not, strictly speaking, be appropriate either. This sample received a low organization score because of the failure in the counterargument and the lack of detail to provide concrete evidence, which is also reflected in the low total word count of 106 words.

Example 4 presents the multiple ideas in a sentence anomaly. In the following example from Student 111, sentence 7 is fairly long and contains multiple ideas. It should have been divided into several sentences.

> **Example 4 (#1-4): Multiple ideas in a sentence** (Student 111)
> ←? [4] Second, we make new communities. ←DET [5] When we change jobs, we meet new job's colleagues. ←? [6] They may give us new surprised. ←? [7] On the other hand, there is view that it isn't beneficial for workers to change jobs often because it is difficult to change jobs and someone may have not good feel for us and thing that "why did you change job? You may have not good personal."

Perhaps the writer was not sure how to present the elements of the counterargument, that is, presenting a different point of view and explaining it. He not only used the conjunction "and" in parallel, but also included a speech line. His discomfort was great because the previous sentence units were relatively short.

Example 5 presents the final anomaly in Category 1, namely convoluted ideas. Student 217 makes the reader feel somewhat uncomfortable because the logic is going in circles at Sentence 5.

> **Example 5 (#1-5): Convoluted ideas** (Student 217)
> ←SUP [4] If working space is bad for them, they can't work smooth. ←? [5] I think a man who feel his working space is bad has to think about change jobs because he has a chance to get more good working space.

One of the reasons why the sentence appears convoluted is probably that the subject switches constantly from the preceding sentence 4 to sentence 5, which means that the theme is not clearly defined. This would make it difficult for the reader to determine where to focus his or her attention. Furthermore, the alternating presentation of working spaces as "good" and "bad" from different perspectives also gives an impression of uneasiness.

As has been demonstrated thus far, the Category 1 anomalies have a wider scope than the other categories. Therefore, multiple factors are possible, with varying degrees of severity. This may have led to the occurrence of Category 1 anomalies over a wide range of score levels.

4.2.2.3.2 Category 2: Deviation From the Topic. The writing samples that presented Category 2 anomalies, indicating the presence of sentence unit with deviation from the topic, scored relatively low. After all, the presence of off-topic sentence units seems to be judged as detrimental to the coherence of the paragraph. This category applies to Students 208 and 227, both writing samples from the intervention group, but both have relatively high total word counts, especially the one from Student 208, which was the third longest sample of all. The positive challenge of trying to include a lot of content may have contributed to the inclusion of unsuitable sentence units. Example 6 is from Student 208.

> **Example 6: Deviation from the topic** (Student 208)
> ←SUP [6] Second, changing company let you learn many things. ←DET [7] Of course, when you change jobs often, you can experience so many new things. ← ? [8] Then, after you get these skills, you can have chance to do things you want like building a new company. TP [1] ←CA [9] Some people say you can get less money when you change jobs many times.

In sentence 8, the topic has suddenly shifted from changing jobs to establishing a new company, which is not the main point of the story. It could be an extension of the idea that one can do what one wants to do and may not be entirely irrelevant, but the subsequent sentence shifts to a different issue and abandons the topic alone. Therefore, it is a deviation from the topic.

4.2.2.3.3 Category 3: Redundant Ideas. Category 3 is an anomaly caused by redundancy; it mostly applied to low-scoring samples, except for one case. A close examination of the sentence of the exceptional case from Student 103 (Example 7) revealed that the last sentence right before the concluding sentence was judged to be redundant, suggesting that the student writer may have restated the same content, perhaps as a summary, and added a sentence at the end to increase the word count to meet the requirement. Without the sentence in question, the total word count would have been reduced by 13 to 93 words, which would not have satisfied the requirement to write at least 100 words.

> **Example 7: Redundant ideas** (Student 103)
> ←DET [6] So, we should experience a lot of jobs by changing jobs often. ←? [7] Second, we take chance to challenge by changing job. ←? [8] Changing jobs mean new challenge. ←DET [9] Their experiences are important for future. ←? [10] I think we should experience a lot of challenge by changing jobs often. ←CON [11] For these reasons, I think it is beneficial for workers to change jobs.

4.2.2.3.4 Category 4: Lack of Organizational Constituent(s). Category 4 includes anomalies related to the lack of an organizational constituent(s). Six out of the seven instances were caused by problems located in the topic sentence or the concluding sentence. The samples from Students 112, 201, and 220 had problems identified in the topic sentence. None of these students clearly stated their opinion position in the topic sentence; rather, they started in an ambiguous way. The subsequent supporting and concluding sentences recovered and corrected the delay, but uncertainty at the start means that it takes time for the reader to be convinced of the direction of the statement. The samples from Students 121, 111, 205, and 124 have problems with the concluding sentence. The concluding sentence may be carelessly missing because the element appears at the end of the passage.

The following two examples are excerpts of anomaly in a topic sentence (Student 201) and a concluding sentence (Student 124). Although it is possible to start the first sentence of a paragraph as an introductory sentence with a question to attract the reader's attention, sentence 1 of Example 8 does not relate to the required topic. Sentence 3 describes a job change, but it is too weak to provide a clear claim as a topic sentence.

Example 8: Lack of organization – topic sentence (Student 201)
? [1] Should contemporary people keep working until they retire? ←DET [6] The answer is No. ←DET [3] To change job is more beneficial due to below reasons.

In Example 9, the passage ends with rebuttal support at sentence 10 and lacks a concluding sentence to summarize the passage. The writer may have felt relieved after finishing the counterargument as well as the rebuttal and its support, but because sentence 10 cannot serve as a conclusion for the entire paragraph, a concluding sentence would still be necessary.

Example 9: Lack of organization – concluding sentence (Student 124)
TP [1] I think it isn't benefit for workers to change jobs often. *(omitted in the middle)* TP [1]←CA [8] Certainly by change jobs they often can experience and understand a lot of things such as processes, rules, methods, structures. ← REB [9] But especially in the case they have family, they must earn much money for their family. ←DET [10] I think it is important for workers to have high and stable income. ←?[11] NO CONCLUSION

4.2.2.3.5 Category 5: Misunderstanding the Task/Task Incompletion. There was only on Category 5 anomaly (a problem with task completion). The problem in Example 10 is located in the topic sentence. This anomaly type is attributed to a student's misunderstanding of the task, which is not so common for the target students in the present study. In this case, the student started off on the wrong foot and the discussion ended with a different conclusion from that required by the task. For this anomaly type, incompleteness of the task is identified by checking both the topic sentence and the concluding sentence as well as the contents of the intermediate steps.

Example 10: Misunderstanding the Task (Student 220)
? [1] I agree with the topic that whether it is beneficial for workers to change jobs often or not. ←? [2] I think that when people change jobs, there are two case. *(omitted in the middle)* ←CON [13] Above all these reasons, I think that workers should choose jobs that they fit.

This type of anomaly is not necessarily a problem of composition, as it is primarily about content. In the sense of “consistently misinterpreting the task,” this anomaly type may be different from other anomaly types, which undermines the coherence of other paragraphs. However, of course, writing with these anomalies receive lower scores.

4.2.2.3.6 Category 6: Inappropriate Support Due to Insufficient Knowledge on the Subject – Relevance of Support. Category 6, along with Category 1, had the most frequent anomalies. This category is related to the relevance of support and defined as inappropriate support due to insufficient knowledge on the subject. The number of Category 6 anomalies is expected to be relatively high because of the high number of sentence units providing support and details. This type of anomaly was observed regardless of the score level. However, the seriousness and frequency varied from sample to sample, and the scores seemed to vary accordingly. In Example 11 from Student 113, because sentence 2 states that changing jobs gives workers the opportunity to gain more experience, sentence 3 should provide an explanation or specific example of this. However, sentence 3 describes the importance of connecting with people and their experiences, and this is not at all specific to job changes, but only refers to relationships in general. In other words, sentence 3 should serve as a minor support that specifically addresses sentence 2, but it does not succeed due to insufficient demonstration of knowledge specific to the job change.

> **Example 11: Inappropriate Support** (Student 113)
> TP [1] In my opinion, it is beneficial for workers to change their job often. ← SUP [2] First, changing job gives us a lot of experiences. ←? [3] Almost all jobs need the connection of people, so workers can make strong bonds with others. ←DET [4] It also improves our communication skills.

4.2.2.3.7 Category 8: Insufficient Rebuttal. Category 8 corresponds to rate ideas and rhetoric in the study by Cumming et al. (2001, 2002). In the present study, the judgment was conducted only on the issue of the success or failure of incorporating the rebuttal into the paragraph in terms of identifying an anomaly in the sentence. Hence, the data are exactly the same as the rebuttal of the organizational element in Table 4.19. In Example 12, Student 219 is in favor of a worker changing job often because it would improve his/her career. Sentence 6 is a counterargument where he presents his concern about the public criticism that job changers receive, including that they lack patience. In sentence 7, as a rebuttal, the writer needs to present an argument that negates or resolves the concern, but the writer does not mention it and makes a rebuttal that misses the point that you cannot grow if you do not change jobs.

> **Example 12: Insufficient Rebuttal** (Student 219)
> TP [1] I believe for workers changing jobs is a way to improve their own careers. ←? [2] No one knows if the jobs they want to do is the right one for the. TP [1] ←SUP [3] Even if they think their current job is not a good fit. They can find a job that really suits them by changing jobs. ←DET [4] If the job is not for workers. Then they should try a new stage. ←DET [5] And, workers can prepare themselves with a lot of social experience before going to the next step. TP [1] ←CA [6] Some people say they have no patience. ←? [7] However. I don't think that continuing to work at a job that doesn't suit will help they improve their skills.

One point that needs to be made regarding anomalies in the rebuttal in all the writing samples in the Transfer C task is that there were 12 writing samples in total, five in the control group and seven in the intervention group, that did not include the approach to combine a counterargument and rebuttal from the outset. Although the prompt recommended using this approach, some students decided not to include it. It is possible that some students who experienced difficulties during the process writing revision work avoided this approach when completing the transfer task. In fact, the difficulty in using rebuttal was frequently observed in the responses to the questionnaires. There were writing samples that did not include this approach at each score level: three at level 5 (two control and one intervention), four at level 4 (two control and two intervention), five at level 3 (one control and three intervention), and one at level 2 (in the intervention group). The fact that some of the writing samples without this approach received a score level of 5 or 4 indicates that their decision sometimes worked.

4.2.2.3.8 Category 11: Incomprehensible Sentence – Related to Language Use.

Lastly, Category 11 is a language-related issue that causes stagnation due to incomprehensible sentence. There were 12 writing samples with a Category 11 anomaly, of which three scored level 4 or higher and eight scored level 3 or lower (Table 4.18). At the sentence unit level, there were 15 instances, of which four scored above level 4 and 11 scored level 3. The frequency of the lower scoring groups is by far the higher. Frequency is not necessarily the only factor that reduces the score; the severity of the problem and the location of its occurrence also contribute, so it is difficult to generalize. However, it is certainly a factor that pushes scores down. A closer look at the three writing samples from Students 111, 213, and 117, which scored highly even with a Category 11 anomaly, revealed the following factors. For Student 111, the anomalous unit in question is at the end of the structure, so it does not affect other sentences and does not bother the reader much. For Student 213, there is no other anomalous unit, and therefore, the meaning of the sentence can be understood. For Student 117, the sentence unit in question can be predicted and complemented by the reader to some extent from the sentence before and after, and furthermore, and the structure in the overall tree-view is well-balanced. On the other hand, Example 13 from Student 207 has only one anomaly of this type in the whole sample. However, the unintelligible sentence 3, which is the beginning of the presentation of reasons, provided a bad start. It stopped the flow of the passage at the very first major support to presumably increase the severity of the problem and resulted in a lower score.

Example 13: Incomprehensible sentence (Student 207)
←DET [2] I have two reasons. ←? [3] First, it is more difficult for workers to continue their job until they stop working than past. ←DET [4] Hence, they must look for next job if you can't continue your job.

4.2.2.4 Analysis of the Overall Structure/Shape of Annotated Tree Diagrams.

In this section, some representative examples of tree diagrams created by tagging the students' writing with the annotation tool are examined to explore the relationship between the integrated organization and content scores and tree diagrams in the transfer task. Furthermore, the structure or shape of all tree diagrams for the transfer task in the two groups was investigated to examine whether there were differences in the features of the tree diagrams between the control and intervention groups. The first half of these two analyses would indicate the effectiveness of schematic information as an analytical tool for instructors and could be considered to provide evidence to address RQ 3. The second half of the analysis, as in the previous analyses, qualitatively examines the retention effect, which would also contribute to RQ 3 although the control and intervention groups were quantitatively almost the same in this respect.

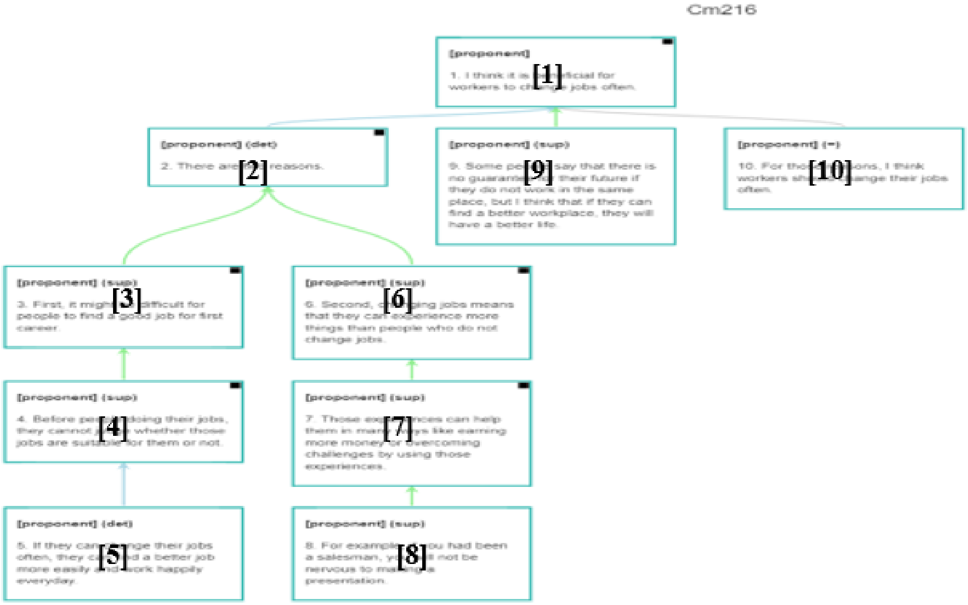
4.2.2.4.1 Examples of Tree Diagrams of Samples That Received High Scores. Figures 4.9 through 4.11 show the tree annotation diagrams and sentence transcriptions of the three writing samples that received high organization and content scores. For supplementary information, the 6-point language use, vocabulary, and mechanics scores are also provided. The figures are assigned a number corresponding to the sentence number. The sentence units are preceded by arrows indicating the connection between the units; the starting point is the source sentence, the ending point is the target sentence, and the organizational element is written in abbreviated form. Anomalous units are indicated by a question mark in a bold box in a diagram and as an underline in a sentence.

In Figures 4.9–4.11, the high-scoring tree diagrams appear to have some features in common: they are concise, balanced, and well-formed.

Figure 4.9 presents a diagram generated from the sample that received the highest score. The writing sample has the most cost-effective structure in that the minimum required elements are well placed following the standard format of argumentation. Specifically, it has the following organizational elements: one topic sentence (sentence 1) at the beginning to clearly claim her argument, a concluding sentence at the end to restate the claim, a sentence unit serving as an advance notice in the listings (sentence 2), and two major supports for topic sentence to give specific reasons (sentences 3 and 6) accompanied by two minor supports each (sentences 4 and 5 and 7 and 8) to give more details for reasoning. Furthermore, sentence 9 appropriately and succinctly incorporates a counterargument, a point of view different from her own, and a rebuttal to negate it immediately afterwards in a single sentence by explaining the view that some people in the public may insist on the stability of continuing in the same job but finding a better new job can lead to a happier life. The whole passage appears coherent throughout as the writing follows the format according to the reader's expectations, reducing the burden on the reader and facilitating understanding. Moreover, this consistency resulted in no anomaly identified in the passage.

Figure 4.9

Balanced Tree Diagram for Student 216



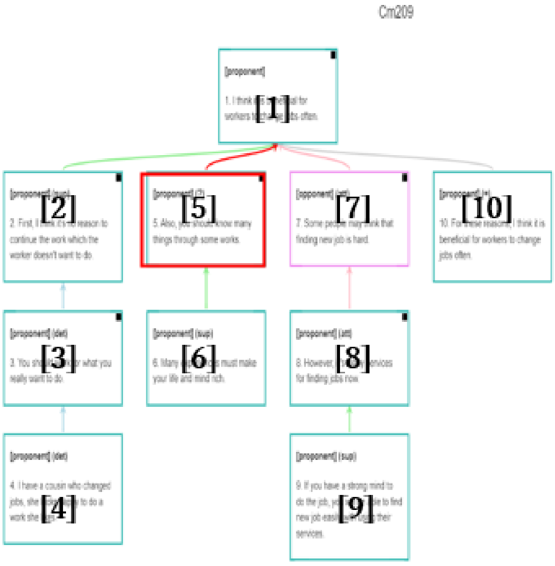
[1] I think it is beneficial for workers to change jobs often. ←DET[2] There are two reasons. ←SUP[3] First, it might be difficult for people to find a good job for first career. ←DET[4] Before people doing their jobs, they cannot judge whether those jobs are suitable for them or not. ←DET[5] If they can change their jobs often, they can find a better job more easily and work happily every day. [2]←SUP [6] Second, changing jobs means that they can experience more things than people who do not change jobs.←DET[7] Those experiences can help them in many ways like earning more money or overcoming challenges by using those experiences. ←DET[8] For example, if you had been a sales man, you will not be nervous to making a presentation. [1]←SUP[9] Some people say that there is no guarantee for their future if they do not work in the same place, but I think that if they can find a better workplace they will have a better life. [1] ←=[10] For these reasons, I think workers should change their jobs often. (167 words)

Note. This example is from the intervention group; the scores were 6 for organization and 6 for content. The numbers in the boxes indicate the order of the sentences.

Figure 4.10, one of the highest scored writing, also shows a balanced structure. This example is characterized by the presence of rebuttal support at sentence 9 in addition to a counterargument–rebuttal pair at sentences 7 and 8, with this one continuous group of sentence units (test units of sentence 7–9) serving as support for a persuasive argument by introducing the existence of an agency to help people change jobs. On the other hand, sentence 5 was identified as an irrelevant support, or an anomaly, which was not well connected to sentence 4, and a reader would be confused by the lack of clarity about the role and direction of sentence 5 in the passage. However, the fact that the immediately following sentence 6 was well complemented, and the subsequent series of counterarguments and rebuttal argumentation was effective, suggests that the problems with sentence 5 were not fatal after all, and the evaluation was favorable.

Figure 4.10

Balanced Tree Diagram for Student 209



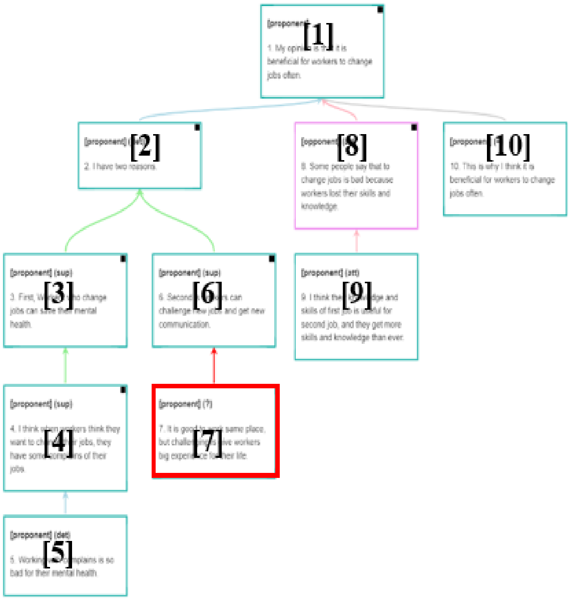
[1] I think it is beneficial for workers to change jobs often. ←SUP[2] First, I think it's no reason to continue the work which the worker doesn't want to do. ←DET[3] You should work for what you really want to do. ←DET[4] I have a cousin who changed jobs she looks happy to do a work she likes. [1] ←?[5] Also, you should know many things through some works. ←DET[6] Many experiences must make your life and mind rich. [1] ←CA[7] Some people may think that finding new job is hard. ← REB[8] However, it's many services for finding jobs now. ←SUP[9] If you have a strong mind to do the job, you will be able to find new job easily with using their services. ←= [10] For these reasons, I think it is beneficial for workers to change jobs often. (127 words)

Note. This sample is from the intervention group; the scores were 6 for organization and 5 for content. The bold box in the tree diagram and the underlined sentence represent an anomalous sentence unit.

Figure 4.11 is also derived from one of the highest scored writing with one Category 1 anomaly identified at sentence 7. The structure of this example is basically the same as Figure 4.9, except for the inappropriateness of the minor support sentence 7 against the major support sentence 6. Because sentence 6 described "getting new communication" by changing jobs, sentence 7 should have given further description or a specific example as detailed support. However, sentence 7 did not logically support sentence 6 sufficiently. Due to the presence of an anomaly in sentence 7, the sample received a content score of 5. Unlike the case of sentence 5 in Figure 4.10, the inappropriateness of sentence 7 was not complemented and moved to the next argument stage in sentence 8. In addition, this writing had several language use errors in the combination of be verbs and general verbs in sentences 6 and 7. These errors did not impede understanding the meaning, so they did not severely affect the content score.

Figure 4.11

Balanced Tree Diagram for Student 122


[proponent]
[1]
1. My opini is that it is beneficial for workers to change jobs often.
[proponent]
[2]
2. I have two reasons.
[opponent]
[8]
8. Some people say that to change jobs is bad because workers lost their skills and knowledge.
[proponent]
[10]
10. This is why I think it is beneficial for workers to change jobs often.
[proponent] (sup)
[3]
3. First, W o change jobs can save their mental health.
[proponent] (sup)
[6]
6. Second, s can challenge new jobs and get new communication.
[proponent] (att)
[9]
9. I think the dge and skills of first job is useful for second job, and they get more skills and knowledge than ever.
[proponent] (sup)
[4]
4. I think when workers think they want to ch jobs, they have some compl of their jobs.
[proponent] (?)
[7]
7. It is good to work same place, but challenging is give workers big experience for their life.
[proponent] (det)
[5]
5. Working plains is so bad for their mental health.

[1] My opinion is that it is beneficial for workers to change jobs often.←DET[2] I have two reasons. ←SUP[3] First, workers who change jobs can save their mental health. ←DET[4] I think when workers want to change their jobs, they have some complains of their jobs. ← DET[5] Working with complains is so bad for their mental health. [2]←SUP[6] Second, is workers can challenge and get new communication. ←? [7] It is good to work same place but challenging is give workers big experience for their life. [1]←CA[8] Some people say that to change jobs is bad because workers lost their skills and knowledge. ← REB[9] I think their knowledge and skills of first job is useful for second job and they get more skills and knowledge than ever. ←=[10] This is why I think it is beneficial for workers to change jobs often. (136 words)

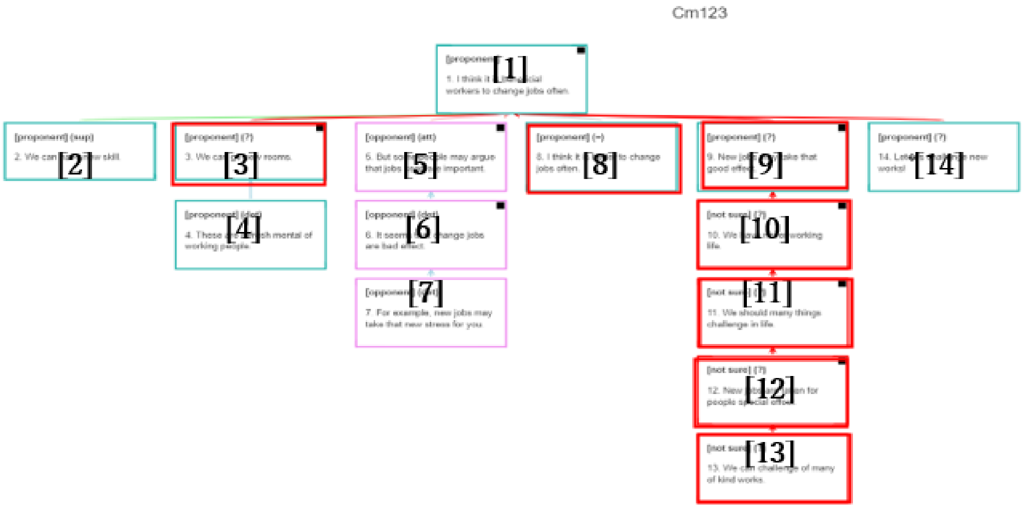
Note. This sample is from the control group; the scores were 6 for organization and 5 for content.

4.2.2.4.2 Examples of Tree Diagrams of Samples That Received Low Scores.

Figures 4.12–4.14 are examples of low-scoring structures of the three representative structural shapes: horizontally wide, unbalanced, and vertically long. Figure 4.12 is characterized by the presence of multiple sentence units in the first layer, giving it a horizontally wide appearance. Furthermore, the lengths of the bead-like units, which consist of details, also varied, which indicates insufficient and unbalanced details. When reviewing the actual sentence content, sentence 3 is abruptly read as “We can go new rooms,” which seems to be a metaphorical expression. However, it is unclear whether it is an additional description of sentence 2 or a new sentence to support the topic sentence. Sentence units 5–7 can be interpreted as a counterargument and rebuttal. Sentence 8 recapitulates sentence 1, the topic sentence. This is probably because the series of arguments had been discontinued after sentence 7. The writer probably could not come up with an appropriate connector and could only come up with this kind of approach to develop a new argument, but it resulted in a short passage that divided the whole sample into two parts. This definitely leads to a lack of coherence throughout the passage at the macro level. At the micro level, the anomalous sentence units 9–13 lacks a sense of sequence and does not work as supporting sentences appropriately, which led to an incoherent paragraph.

Figure 4.12

Horizontally Wide Tree Diagram for Student 123



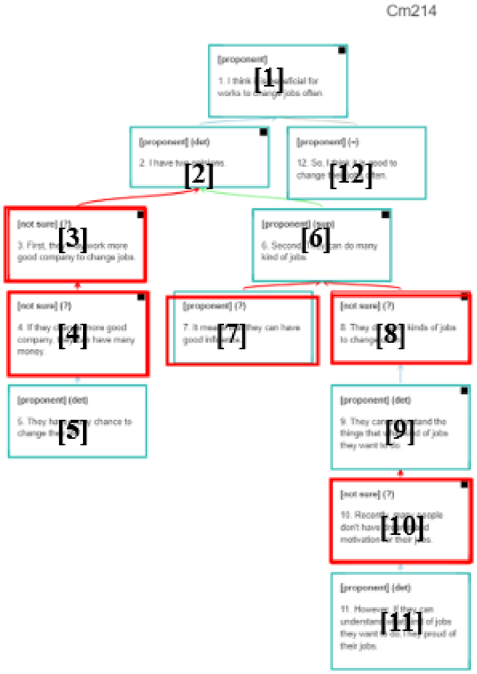
TP [1] I think it is beneficial workers to change jobs often. ←SUP[2] We can have new skill. [1] **←?** [3] We can go new rooms. ←DET[4] These are refresh mental for working people. [1] ←CA[5] But some people may argue that jobs keep are important. ←DET[6] It seems that change jobs are bad effect. ←DET[#7] For example, new jobs may take that new stress for you. [1]←? [8] I think it is better to change jobs often. [1] ←? [9] New jobs may take that good effect. ←? [10] We have new working life. ←? [11] We should many things challenge in life. ←? [12] New jobs are taken for people special effect. ←? [13] We can challenge of many of kind works.

Note. This sample is from the control group; the scores were 2 for organization and 2 for content.

Figure 4.13 shows an unbalanced structure with scattered anomalies, with irregular branching and variations in the number of details. There is a problem of expression in sentence 3, where it is ambiguous whether the “to” infinitive is an object or a result, which makes the sentence sound illogical. However, the immediately following sentence 4 was judged to be a logically incorrect statement because the job change does not necessarily lead to a salary increase, nor was it explained. Likewise, sentence 7 fails to serve as a support for sentence 6, and sentence 8 convolutes the discourse by going backwards and not moving forward. Furthermore, sentence 10, even though it is at the end of the paragraph, brings up the new topic of “motivation” ignoring the previous discussion, resulting in a lack of coherence.

Figure 4.13

An Unbalanced Tree Diagram for Student 214



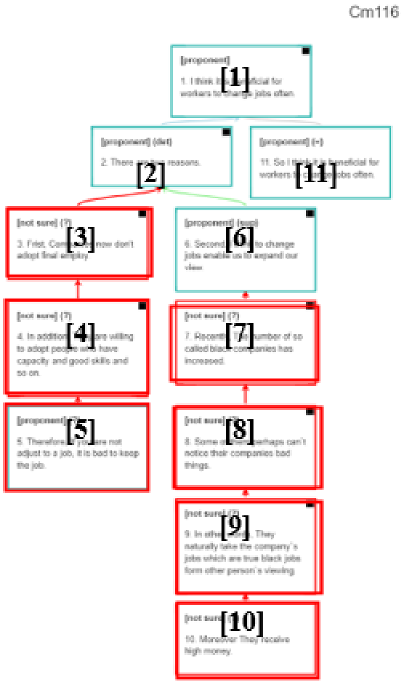
[1] I think it is beneficial for workers to change jobs often. ←DET[2] I have two reasons. ←? [3] First, they may work more good company to change jobs. ←? [4] If they change more good company, they can have many money. ←DET[5] They have many chances to change their life [2] ← SUP[6] Second, they can do many kind of jobs. ← ? [7] It means that they can have good influence. [6] ←? [8] They do many kinds of jobs to change often. ←DET[9] They can understand the things that what kind of jobs they can do. ← ? [10] Recently, many people don't have dreams and motivation for their jobs. ←DET[11] However, if they can understand what kind of jobs they want to do, they proud of their jobs. ← = [12] So, I think it is good to change their jobs often. (123 words)

Note. This samples if from the intervention group; the scores were 2 for organization and 4 for content.

Figure 4.14 shows an elongated vertical tree diagram, with several details following one supporting sentence, and information may be added like a narrative, making it less concise. First, sentence units 3 and 4 explain the employer's point of view, with sentence 3 discussing lifetime employment and sentence 4 discussing skill-based hiring. These are factors that encourage workers to change jobs, but it is not well explained and not well connected to job changes on the part of workers, which resulted in being identified as anomalies. Next, sentences 7–10 were identified as anomalies of improper order. Some students use this type of approach to state "the background first and then gradually move on to the point they want to make" (Matsumura & Sakamoto, p. 42), an approach that sometimes results in difficulty for a reader to understand the connection between these statements and the claim. In addition to the roundabout expressions, there is no counterargument or rebuttal in the passage, which makes the argument monotonous and lacking in persuasiveness, leading to the low content score.

Figure 4.14

A Vertically Long Tree Diagram for Student 116



[1] I think it is beneficial for workers to change hobs often. ←DET[2] There are two reasons. ←? [3] First, Companies now don't adopt final employ. ← ? [4] Ind addition, they are willing to adopt people who have capacity and good skills and so on. ← ? [5] Therefore, if you are not adjust to a job, it is bad to keep the job. [2] ←SUP[6] Second, I think to change jobs enable us to expand our view. ← ? [7] Recently the number of so called black companies has increased. ← ? [8] Some of them perhaps can't notice their companies bad things. ← ? [9] In other words, they naturally take the company's jobs which are true black jobs from other person's viewing. ← ? [10] Moreover, they receive high money. ← = [11] So I think it is beneficial for workers to change jobs often. (122 words)

Note. The sample is from the control group; the scores were 6 for organization, 3 for content, 2 for language use, 2 for vocabulary, and 2 for mechanics.

Based the six examples of tree diagrams, the shape of the diagram and the number of anomalies as well as their positions are related to some extent to the organization and content scores. These factors are subtly interrelated and not always easy to generalize succinctly, but it seems certain that the analysis using the annotation tool helped to lead to these findings. The following section presents the results of an extended examination of how the structure of the tree diagram differs between the control and intervention groups on the transfer task.

4.2.2.5 Comparison of the Tree Diagrams Between the Control and Intervention Groups. Table 4.21 summarizes the results of comparing the shape of the tree diagrams for the transfer task between the control and intervention groups by classifying them by the four types of structure presented in Section 4.2.2.4.

Table 4.21

Frequencies and Percentages of Types of Shape of Tree Diagrams on the Transfer Task in the Control and Intervention Groups

Structural type	Control group (*n* = 22)		Intervention group (*n* = 23)	
Balanced	11 (50.0%)	50%	16 (70.0%)	70%
Horizontally wide▲	3 (13.6%)	▲50%	2 (8.7%)	▲30%
Unbalanced▲	5 (22.7%)		4 (17.4%)	
Vertically long▲	3 (13.6%)		1 (4.3%)	

Note. Filled triangles indicate poorly formed diagrams.

Overall, 70% (16/23) of the transfer tasks in the intervention group had a balanced annotated tree diagram shape, compared with 50% (11/22) for the control group. This indicates that the intervention group was superior in the transfer task in terms of structural type of organization of the passage—in other words, the number of each organizational components, their placement, and their links. Given that a well-balanced sentence is difficult to form without good logical connections, it can be considered an indication of good coherence. It should also be noted that there was only one vertically long tree diagram in the intervention group, compared with three in the control group. A vertically long form generally consists of only one support or indicates writing in the narrative form, which is not a very effective approach in English writing, especially as an argumentative passage. This finding suggests that the intervention group was able to compose paragraphs in accordance with coherence and with an approach appropriate for argumentative sentences. Furthermore, it is significant that this is the result of a transfer task, a test of retention of experience and perceptions gained from previous activities. The intervention group received schematized feedback at least twice in the rewriting work, and it is possible that they either consciously or unconsciously aimed at the “ideal form” of a paragraph. In fact, the questionnaire results showed that many students in the intervention group were interested in the novelty and clarity of the tree diagram, and some students even drew a tree diagram to work on during the planning process. This may indicate that they perceived the paragraph as a graphical image.

Although it is difficult to generalize these findings due to the small sample size, it may be concluded that the effectiveness of the schematized feedback on the intervention group was observed qualitatively, if only partially, whereas the scores on the rating scale did not necessarily show significant differences.

4.2.3 Summary of Study 3-1 (Qualitative Results with Annotation Diagrams)

The purpose of Study 3-1 was to explore the effectiveness of schematized feedback in the transfer task through the analysis of information obtained by the annotation tool to address RQ 3(3). To this end, a coding scheme of 11 anomaly categories was prepared by referring to the descriptors in the ESL CP (Jacobs et al., 1981) used in the quantitative and the classification of raters’ judgment strategies suggested by Cumming et al. (2001, 2002) and Barkaoui (2007). Subsequent analyses were conducted in accordance with these coding schemes.

First, the frequencies for each organizational element and the anomaly location in the passage were investigated to determine the overall patterns of anomaly appearances. One hundred instances of anomalies were identified in 45 writing samples. There was no significant overall difference between the control and intervention groups in terms of the anomaly location. However, the intervention group had zero counterargument anomalies, whereas the control group had four.

Second, the relationship between the anomaly location in terms of the organizational element was explored. Although there was no apparent overall relationship between the scores and the anomaly categories, the writing samples in which an anomaly was found in the topic sentence and the counterargument tended to have low scores. However, it may be difficult to generalize this result due to the small number of applicable cases.

Third, the frequencies of anomaly categories in the control and intervention groups were investigated. Of the total 100 anomalies, 59 were in the control group and 41 in the intervention group. Among them, Category 1, poor reasoning, logic, and topic development, had the most anomalies (24 in the control group and 10 in the intervention group), followed by Category 6, irrelevant support (15 in the control group and 7 in the intervention group). These results showed that there were more anomalies that hinder coherence in the control group than in the intervention group. Given the moderate negative correlation between the content and organization scores and the frequency of anomalies, the number of anomalies have some negative impact on the scores.

In the next part of the analysis, the relationship between the anomaly category and the scores was presented along with excerpts from an actual writing sample of students. As the notes to the actual examples indicate, the relationship between the anomaly category and score cannot be simply explained. Although the relationship depends on the type, it is deeply related to the severity of the problem, the anomaly location, or whether it is successfully corrected in the sentence unit that follows. However, several trends were identified, including lower scores for Category 2 (deviation from the topic), Category 3 (redundancy), and Category 4 (irrelevant support) anomalies, with some exceptions.

Lastly, the structures or shapes of the tree diagrams generated with TIARA annotation tool were examined. The diagrams were classified into four types according to the general structural features: balanced, horizontally wide, unbalanced, and vertically long. The latter three represent poorly formed diagrams while the former represents a well-formed diagram. The percentage of the three types of poorly formed tree diagrams was higher in the control group (50%) than in the intervention group

(30%). This finding suggests that the intervention group presented more writing samples with a balanced tree diagram, leading to higher organization scores compared with the control group.

4.2.4 Qualitative Analyses Results for Study 3-2: RQ 3(4)

4.2.4.1 Students' Behaviors During Rewriting Based on the Responses to a Questionnaire (Control Versus Intervention). As described in Chapter 3, the first and second paper-based questionnaires were administered immediately after the post-tasks. The questionnaires included two questions, which were written in Japanese and translated into English by the author below.

1. What did you fix or modify during the revision process?
2. What did you think about the feedback from the teacher? Please write your candid opinions.

A standard thematic analysis was conducted by following the eight steps described by Takagi (2021). The themes obtained through the thematic analysis were summarized as a storyline to describe a series of students' behaviors in the process of rewriting (Creswell & Creswell Báez, 2021), which would hopefully contribute to reveal the nature of students' revision work with teacher feedback in the EFL process writing.

4.2.4.1.1 What the Students Revised in the Second Draft With Feedback. This section presents the results of the open-ended questionnaire administered to the control and intervention groups after the rewriting task of either Topic A or B. In the questionnaire, the participants were asked to describe what they had revised in the five analytic criteria: content, organization, vocabulary, language use, and mechanics. For the second question, the students were asked to comments on the teacher feedback provided prior to the revising task. In this section, student responses regarding the modifications are analyzed first. As discussed in Chapter 3, a thematic analysis was conducted on the data from 36 participants (18 from the control group and 18 from the intervention group). Note that content and organization criteria were combined and treated as one category.

Table 4.22 shows the results of the revisions reported by participants that they declared they had made during the rewriting work in the post-tasks. A total of 173 revisions were reported, 78 from the control group and 95 from the intervention group. The difference between the groups was clearly due to changes in content and organization: 21 for the control group and 40 for the intervention group. In other words, there was relatively little difference between the groups in the frequency of student-reported modifications for the other criteria. It should be noted that the frequencies of content and organization, language use, and vocabulary changes in the control group were very similar, 21, 25, and 23, respectively. On the other hand, the intervention group made more revisions in the content and organization category, 40, than the language use and vocabulary categories, 20 and 25, respectively.

Table 4.22

Frequencies of the Revisions Reported by Participants on the Writing Skill Criteria

Writing skill criteria	Control (*n* = 18)	Intervention (*n* = 18)	Total (*n* = 36)
Content and organization	21	40	61
Language use	25	20	45
Vocabulary	23	25	48
Mechanics	9	10	19
Total responses	78	95	173

To examine the difference between the two groups more precisely, a detailed analysis of the students' revision behaviors in the content and organization category was necessary. The next section focuses on this category to analyze the students' comments in more detail based on the thematic analysis.

Table 4.23 summaries the four key themes for the content and organization category drawn from the 61 responses: (a) addition/supplementation (30 responses), (b) addition and revision/alteration (19 responses), (c) rearrangement of sentences (9 responses), and (d) deletion (3 responses). There are several items whose frequency is small and difficult to compare, so the discussion focuses on items whose frequency is relatively large and thus can be compared between groups.

The first point to note when comparing the control and intervention groups is the difference in the frequency of responses. The control group mainly provided responses in three of the eight codes, while the intervention responded to all eight codes. The corrections reported by the intervention group outnumbered those of the control group in both the total frequency and the types, although the report of corrections did not always match the actual corrections. It can be inferred that the respondents in the intervention group were more consciously making revisions in terms of the content and organization criteria, so they recognized and remembered them with certainty and provided more details. Another point to note is that the largest difference in frequency of responses between the two groups was in (b) revision/alteration, with five for the control group and 14 for the intervention group. Theme (b) is a replacement of elements or content of the original text, and should require more major changes than the addition of elements in theme (a). This suggests that the intervention group may have attempted more substantial modifications.

Table 4.23

Theme and Code Frequencies: Reported Revisions Made for the Content and Organization Criteria

	Theme	Code	Control (*n* = 18)	Intervention (*n* = 18)	Total (*n* = 36)
(a)	Addition/ supplementation (30)	Addition and/or supplementation of words and/or phrases	12	15	**27**
		Addition of missing component sentence(s)	0	3	**3**
		Subtotal	12	18	**30**
(b)	Revision/alteration (19)	Modification of conjunctive expressions	3	7	**10**
		Change or revision of content (reasons)	2	6	**8**
		Change to pronoun(s)	0	1	**1**
		Subtotal	5	14	**19**
(c)	Rearrangement of sentences (9)	Sentence reordering	2	4	**6**
		Sentence splitting and merging	2	1	**3**
		Subtotal	4	5	**9**
(d)	Deletion (3)	Deletion of redundant part(s)	0	3	**3**
	Total responses		**21**	**40**	**61**

Note. The frequency of responses exceeds the number of participants because some respondents may have given more than one response. The shaded cells indicate no applicable response.

There were also differences between the control and intervention groups regarding the content of the descriptions. The intervention group tended to be more detailed and specific in their descriptions. The following examples are responses from both groups: Examples 1–6 are from the control group, and Examples 7–12 are from the intervention group. All original responses written in Japanese were translated into English by the author. The translation is followed by the original Japanese in round brackets. Italics indicate the code presented in Table 4.23. Moreover, the pre- and post-task content and organization scores for the task immediately preceding each comment description are listed in square brackets for reference.

In the excerpts from the control group, some students provided a good description of what they had and had not revised for specific reasons. These statements are underlined. The specific statements are as follows: in Example 1, "The structure was not particularly changed"; in Example 3 "I didn't make any major changes"; in Example 4, "I made no significant changes from the previous one"; and in Example 5, "The overall structure was not changed significantly." These modifications were likely partial and did not involve corrections to the entire paragraph. Of the five underlined examples, there was no improvement in the content and organization scores for Examples 1–4, although Example 5 did show improvement. This may be an indication that partial modifications had little impact on the content and organization scores. For Example 5, the student stated "the overall structure was not changed significantly," but there was a 2- or 3-point increase in the content and organization scores. In other parts of the description, it says that the content of the rebuttal was enriched, which suggests that a better approach in the counterargument and rebuttal contributed to the improvement in the rating. Finally, in Example 6, sentence reordering may have prompted the organization score.

- Responses from the control group

Example 1 (Student 102) *Addition/supplementation of words or phrases*:

> I added a supplementary explanation because my opinion was not persuasive enough. The structure was not particularly changed.（自分の意見が説得性に欠けていたため補足説明を追加した。構成は特に変えなかった。）[Content pre-post: 3→3; Organization pre-post: 4→4]

Example 2 (Student 103) *Addition/supplementation of words or phrases*; *modification of conjunctive expressions*: The first reason and supplementary phrases were added. I also added conjunctive phrase. No other change was made particularly.（一つ目

の理由に補足を付け加えた。接続詞を付け加えた。他には特に変えなかった。）[Content pre-post: 2→2; Organization pre-post: 2→2]

Example 3 (Student 104) *Change or revision of content*: I didn't make any major changes, but I've corrected the part of the sentence that was in support of the opposing viewpoint. The abstract part was made more detailed, and made it more readable.（大きな変更はしなかったが、対立の肩を持つような文章になってしまっていたところは修正した。抽象的な部分をより詳しく述べるようにして文を読みやすくした。）[Content pre-post: 3→3; Organization pre-post: 2→2]

Example 4 (Student 116) *Sentence splitting and merging*: I made no significant changes from the previous one. One sentence was split into two.（以前のものと特に大きな変更点はない。1文を2文に変更した。）[Content pre-post: 4→4; Organization: 4→4]

Example 5 (Student 113) *Addition/supplementation of words or phrases*: I included a specific example. I explained the rebuttal part in more detail to make it more persuasive. The overall structure was not changed significantly. 具体例を入れた。説得力を高めるために反論部分をもう少し細かく説明した。全体の構成は大きく変えなかった。）[Content pre-post: 3→6; Organization pre-post: 3→5]

Example 6 (Student 107) *Change or revision of content*; *sentence reordering*: I replaced or added phrases where the reasons were difficult to understand. I changed the order of the sentences and changed the conjunction.（理由が分かりにくかったところの文を入れ替えたり足したりした。順番を変えて接続詞を変えた。）[Content pre-post: 3→3; Organization pre-post: 2→3]

The intervention group paid particular attention to linkage with adjacent sentences. For example, Example 7 says, "I inserted a connecting phrase to indicate clearly that it is an opinion that differs from my own," and Example 8 states, "I checked for contradictions and overlaps with the preceding and following content." These students paid close attention to the preceding and following sentences when inserting or revising supplemental or additional phrases/sentences. Moreover, some responses in the control group showed that each description seemed more detailed by locating the revised part specifically than those of the control group. This could be done more easily based on the schematic feedback, as each text unit was split and numbered. For example, Example 9 specifies that the student reordered the text units by describing with text frequencies, “I sorted 10 text units by content and reordered them into 1-5-4-3-2-6-7-8-9-10.” In Example 10, as indicated by the underlined statements, the student distinguished the two supporting sentences and clearly indicated and corrected the problem. In Example 11, the student concisely explained her correction and how to make them, indicating the location with text frequencies. Finally, Example 12 presents a comment on deletion of redundant part(s), a modification reported only by the intervention group. For this example, both the content and organization scores improved by 2 points. Locating and removing unnecessary sections along with improving the counterargument and rebuttal content can contribute greatly to enhance the content and organization of the passage.

- Responses from the intervention group

Example 7 (Student 204) *Addition/supplementation of words or phrases*; *addition of missing component sentence(s)*; *modification of conjunctive expressions*; *change to pronoun(s)*: I explained the opposing view in detail. <u>I inserted a connecting phrase to indicate clearly that it is an opinion that differs from my own</u>. The conclusion was not present, so a summary of the opinion was added at the end. The repetitive use of the same subject was replaced by proper nouns.（反対意見について詳しく説明した。<u>反対意見であることを示すためにつなぎとなる表現を挿入した</u>。結論が書けていなかったので最後に意見のまとめを追加した。同じ主語が続いていた部分を代名詞に置き換えた。）[Content pre-post: 2→3; Organization pre-post: 2→3]

Example 8 (Student 206) *Change or revision of content (reasons)*: I rethought and revised the rebuttal to the counterargument. Minor support was added to the second supporting sentence. I tried to think logically without rushing. When I came up with new content while writing, I checked for contradictions and overlaps with the preceding and following content.（反論への反駁を再考し修正した。二つ目のサポートセンテンスにマイナーサポートをつけた。焦らずに論理的に考えようとした。新しい内容を思いついた時に前後の内容との矛盾や重複がないかチェックした。）[Content pre-post: 5→6; Organization pre-post: 6→6]

Example 9 (Student 210) *Modification of conjunctive expressions*; *sentence reordering*: The meaning of the sentence immediately before the concluding sentence was not clear, so the connecting element to the previous sentence was modified to make it easier to understand. The order of the contents was changed. I sorted 10 text units by content and reordered them into 1-5-4-3-2-6-7-8-9-10.（コンクルーティングセンテンスの直前の文の意味が通っていなかったので前の文との関連性をわかりやすく修正した 。内容の順番を変えた。内容別に上から一から十までの番号に分け 1-5-4-3-2-6-7-8-9-10 の順番にした。）[Content pre-post: 2→3, Organization pre-post: 2→4]

Example 10 (Student 216) *Change or revision of content (reasons)*; *sentence reordering*: The order of the introductory part was changed. The first and second reasons used the same thing as an example, and there were parts that were tedious, but I was able to change the wording in the second reason and make it clearer. I could make it clearer because I reordered the sentences in the introductory part.（導入部分の順序を変えた 。一つ目と二つ目の理由双方で同じことを例に使っていてくどくなっていた部分があったが、二つ目の理由の部分で言い方を変えてすっきりさせることができた 。導入部分も大幅に順序を変えることができたのが原因だ。）[Content pre-post: 4→6; Organization pre-post: 5→6]

Example 11 (Student 209) *Sentence splitting and merging*: I changed sentences 6 and 7 to make it one sentence.（6と7の文を変えて一文にした。）[Content pre-post: 2→3; Organization pre-post:4→4]

Example 12 (Student 227) *Deletion of redundant part(s)*: Parts that are not relevant to the topic sentence were deleted. I added depth to the opposing views.（主題文に関係のない箇所を消した。反対の意見に深みを持たせた。）[Content pre-post: 2 →4; Organization pre-post: 2→4]

The responses suggest that the control and intervention groups took different approaches to avoid ruining the coherence of the text when making modifications. While the control group tended to make only some of the required modifications to the text, some of the students in the intervention group made more substantial revisions to the structure by paying attention to adjacent text units, specifically by checking the relationship of a given part with the surrounding text.

4.2.4.2 Students' Perceptions of Teacher Feedback. This section presents the results of the analysis of the responses to the second open-ended question on the questionnaire, "What did you think about the feedback from the teacher? Please write your impressions in a straightforward manner." There were 45 respondents, 22 from the control group and 23 from the intervention group. Each of the comments were segmented into units for analyses. A total of 140 units were identified, 71 from the control group and 69 from the intervention group. The resulting units were coded in an exploratory manner.

Before moving on to the results regarding codes and themes, a concrete example (from the control group) is presented to show how segmentation is actually conducted and how counts are made. The comment, originally written in Japanese, was translated into English by the author. The following example was eventually divided into five text segments, which could be categorized into four different codes and themes. A table of classifications is given after the original text in Table 4.24. See Table 4.25 for the code and theme classifications and definitions. Text unit (1) was tagged as (xi) opprotunity to reflect on my writing, text unit (2) tagged as (i) clarity of FB [feedback], text unit (3) as (viii) recognition of errors/mistakes, and text units (4) and (5) as (xiv) critiques for FB. This would count as a total of five text unit extractions.

> (1) I realized something by reading my own writing by referring to the feedback. I thought I needed vocabulary and background knowledge. (2) Also, my teacher's feedback was easy to understand, and (3) there were many points regarding errors she made that I thought were right. (4) My request is that I would like to see what kind of sentences the teacher would write in response to this topic, and (5) I would also like to see an equivalent sentence that would be given about the score as mine. (Student 117)
>
> (1) フィードバックと照らし合わせて自分の文を読むことで気づくことがあった。自分には語彙力と背景知識が必要だと思った。また、(2)先生のフィードバックはわかりやすくて (3) 誤りの 指摘はごもっともだと思う点がたくさんあった。 (4)要望は、先生だったらこのお題に対してどんな文を書くのか見てみたいのと、(5)点数を付けていただいた時にどれぐらいだとこの点数になるのかなと目安の文なども見てみたい。

Table 4.24

A Sample of Classification With the Codes and Themes

Text number	Segmented text	Theme	Code
(1)	I realized something by reading my own writing by referring to the feedback. I thought I needed vocabulary and background knowledge.	III	(x) *Opportunity to reflect on one's own writing*
(2)	Also, my teacher's feedback was easy to understand,	I	(i) *Clarity of FB*
(3)	there were many points regarding errors she made that I thought were right.	III	(viii) *Recognition of errors/mistakes*
(4)	My request is that I would like to see what kind of sentences the teacher would write in response to this topic,	IV	(xiv) *Critiques for FB*
(5)	and I would also like to see an equivalent sentence that would be given a score as mine.	IV	(xiv) *Critiques for FB*

Note. FB, feedback

Table 4.25 shows a summary of the codes and themes extracted through the thematic analysis accompanied by excerpts from the actual student comments. Initially, the codes were extracted from the questionnaire data and further summarized into themes by the author. Some of the codes were collapsed, and the names of the themes were modified through discussions with another coder. Ultimately, a total of 15 codes and four themes based on those codes were identified. The codes are numbered from (i) to (xv), and the themes from (I) to (IV). Generally, one example each from the control and intervention groups is presented in the excerpt column of Table 4.25. These examples are those that the author judged to be representative of the comments in each group. The excerpts from the control group are marked with the abbreviation Cg and those from the intervention group with the abbreviation Ig. In some codes, if there is no example for either group, only one example is given. All responses were written in Japanese originally and were translated into English by the author. The Japanese version of the original data is shown in parentheses immediately after the English translations.

Table 4.25

Summary of the Codes and Themes of Students' Comments on Teachers' Feedback Extracted by Thematic Analysis

Theme	Code	Excerpts
(I) Evaluation of the format and the content of FB	(i) Clarity of FB	**Cg:** "It was easy to work with the tips given on the revision." (書き直す上でのヒントが与えられていたので作業がしやすかった） **Ig:** "When the graphical representation is made, it is easy to see at a glance what is missing in terms of content." (図式化すると内容に足りないものがひと目でわかる）
(I) Evaluation of the format and the content of FB	(ii) Detail and thoroughness of FB	**Cg:** "My teacher provided checks evenly and without bias." (偏りなくチェックを入れてもらった。） **Ig:** "It was carefully checked and spelling errors were noted in red." (丁寧にチェックしてもらい単語の スペルミスには赤で記されていた）
(I) Evaluation of the format and the content of FB	(iii) Analytic evaluation	**Cg:** "I liked the fact that scoring was broken down into structure, grammar, etc., rather than simply being presented with a score." (点数だけ出すのではなく構造や文法などに細かく分けて採点してくださっていたのがとても良かった。）
(I) Evaluation of the format and the content of FB	(iv) Comparison to FB received in the past	**Cg:** "In Junior High and High school, I was never given such detailed feedback when assigned writings during English class."「中高 の時は 英語の時間に英作文を 課されて ここまで詳細なフィードバックをされたことがなかった」; **Ig:** "I have never received such a detailed Feedback before." (今までこんなに丁寧なフィードバックをもらったことがない。）

Theme	Code	Excerpts
(II) Emotions on FB	(v) Emotion (positive)	**Cg:** “It's nice to get detailed feedback that motivated me to work harder.”（丁寧なフィードバックをもらえるともっと頑張ろうと向上心を持つことができる）; **Ig:** “We were grateful that teacher read our writings and tried to understand the nuances that we were trying to convey.” (僕たちの文章を読み伝えようとしているニュアンスを理解してくれた。）
(II) Emotions on FB	(vi) Emotion (mixed)	**Cg:** “I was worried at first by the severe tone of the feedback but reassured to learn that she (teacher) was not angry.”（最初は厳しめのフィードバックの口調に不安を感じたが怒っているわけではないことを知り安心した） **Cg:** “I was not sure about what I was doing, so it was so nice to have someone review and grade my writing for me.”（自信がなかったのでこうしてみていただけて採点してもらえるということがありがたかった） **Ig:** “It feels good to be able to fix something I wasn't happy with.” (納得のいかなかった部分を直せるのは気持ちが良い。）
(II) Emotions on FB	(vii) Emotion (negative)	**Cg: “**I am not sure if the revision made a good sentence or not, and I am worried that it might have the contrary effect.” (修正したことで良い文になったかわからない、逆効果になっていないか不安 **Ig:** “It was difficult to be aware of structure and coherence based on the feedback.”（フィードバックに基づき構成や一貫性を意識するのは難しかった。）

Theme	Code	Excerpts
(III) Perceived effect of FB	(viii) Recognition of errors/mistakes	**Cg:** “I found many parts I could revise, such as vocabulary and conjunctions by referring to the feedback I received.”（フィードバックを見て語彙や接続詞の使いかたなど修正できる部分が沢山あった） **Ig:** “I realized that I had made mistakes and inadequacies in my writing that I had not noticed when I was writing it.”（自分の文章を自分で書き換えると　作成中は　気づかなかった細かいミスや至らない点に気付けた。）
(III) Perceived effect of FB	(ix) Realization of anomalies	**Cg:** “I had to rethink my counterargument and rebuttal all over again because I didn't structure it right.” (反論と反駁の構成が悪かった為一から考え直した） **Ig:** “When you see a diagram, you can see the holes in an argument that you couldn’t see when it was a single sequential passage.”（図式化されるとひとつながりの文章の時には分らなかった主張の穴に気づくことができる。）
(III) Perceived effect of FB	(x) Opportunity to reflect on one’s own writing	**Cg: “**I became conscious not to use the same expressions repeatedly (in transition assignments) since it was pointed out to me on FB before.” (以前にFB で指摘があったのでトピック C（転移課題では）同じ表現を何度も使わないように意識するようになった） **Ig:** “I realized that I am not good at writing with a sense of structure and coherence.” (自分は構成や一貫性を意識して書くのが難しいのだとわかった。)
(III) Perceived effect of FB	(xi) Gaining objective viewpoints	**Cg:** “It gave me a chance to review my writing objectively.” (自分の文章を客観的に見直せるきっかけになった） **Ig:** “The diagrammatic format makes it easy to see how it is read and how it is perceived to mean when read by others.” (図式化されていたのでほかの人が読んだ時どのように読まれているのかどのように意味をとらえているのかがわかりやすい。）

Theme	**Code**	**Excerpts**
(III) Perceived effect of FB	(xii) Benefits for the future	**Cg:** “I learned a lot of necessary knowledge for writing in English.” (英文を書くために必要な知識についてたくさん知ることができた） **Ig:** “I thought I could use it in my future writing.” (今後のライティングにも活かせると思えた。）
(III) Perceived effect of FB	(xiii) Achieving concise modifications	**Ig:** “I feel that the paragraphs are clearer and easier to read myself when I revise them based on the feedback.” （フィードバックを基に修正するとスッキリとしたパラグラフになって自分でも読みやすいと感じる） **Ig:** “I now know which sentences to join together to make a concise sentence.” (どの文を繋げれば簡潔な文になるかわかった。）
(IV) Critiques	(xiv) Critiques for FB	**Cg:** “It would have been easier to understand if there were actual examples of areas of improvement.” (改善部分の実際の例があるとより分かりやすかった） **Ig:** “I felt that I could have learned more if you could have shown me some other ways of expression and ways of writing in the revision part of the English text.” (英文の修正のところで何個か別の書き方や表現方法を示してもらえるとさらに勉強になったかなと感じた。）
(IV) Critiques	(xv) Critiques for assignment	**Cg:** “I felt the connection between what I learned in the textbooks and the writing process was a little weak.” (テキストで学んだこととライティング　作業の相関関係が少し弱いように感じた。） **Ig:** **“**The topic in the assignment is difficult and I can’t think of a specific example.” (課題トピックが難しいので具体例が思い浮かばない。）

Note. Cg, control group; FB, feedback; Ig, intervention group

Theme I, evaluation of the format and the content of feedback, is a summary of codes that describe how the students evaluate teacher feedback. As codes (i) and (ii) show, some found the teacher corrective feedback to be clear, and others found it to be detailed and thorough. Several students gave their impressions of the feedback provided in this study by comparing it with feedback that had been given in their school days. One point that must be noted here is that only code (iv) mentions the feedback sheet with analytic scores based on the rating scale of the ESL CP by Jacobs et al. (1981), while all comments in the other codes and themes in the table concern the corrective feedback. Theme II, emotions on feedback, summarizes the comments about the students' emotions when receiving the feedback. The student responses in this category can be categorized into three codes, positive, mixed, and negative emotions. Positive emotions included a certain frequency of comments of appreciation for the teacher's effort and respect paid to their writing. The experience of receiving detailed writing feedback was new to them, and they seemed simply happy to have their work read by a teacher. Theme III, perceived effect of feedback, summarizes six codes regarding the student's perceived benefits from the feedback. It covered a wide range of aspects, including what helped them to actually revise their writing, what gave them an opportunity to be objective about their writing, and what skills they think useful in their future writing. Finally, theme IV, critiques, has two codes, request for feedback and request for the assignment, where the students commented on what they want further in feedback and what they require in relation to the assignment and class content.

Table 4.26 shows the breakdown of the frequency of text units for the codes and themes extracted in the thematic analysis. A total of 140 text units were identified, 71 in the control group and 69 in the intervention group. As shown in the table, 49 comments were categorized into theme I, 23 into theme II, 54 into theme III, and 14 into theme IV. The cells with no corresponding code, such as code (iii) in the intervention group and code (xiii) in the control group, are shaded in gray.

The total frequencies of comments for the control and intervention groups are nearly equal at 71 and 69, respectively, but in seven, or almost half of the 15 categorical codes, there were some differences in the frequencies of responses between the two groups. Among them, two points are worth noting. The first is the reversal in the frequencies of codes (i) and (ii) in theme I between the groups. The second is the markedly higher number of comments in code (viii), recognition of errors/mistakes, in the control group. At the same time, there are also some codes that are common to both groups, in terms of frequencies and/or content of comments, such as code (iv) in theme I and, in some instances, code (v) in theme II regarding the Emotions toward feedback. Both the similarities and the differences in the frequency of comments and/or contents between the two groups as well as their possible factors contributing are analyzed in further detail, citing actual comments.

Table 4.26

Breakdown of the Frequencies in Codes Extracted by Thematic Analysis

Theme	**Code**	Control	Intervention	**Subtotal**
(I) Evaluation of the format and the content of FB (49)	(i) Clarity of FB	7	20	**27**
	(ii) Detailed and thorough FB	11	2	**13**
	(iii) Scores by analytical evaluation	4	0	**4**
	(iv) Comparison to FB received in the past	3	2	**5**
		(25)	**(24)**	**(49)**
(II) Emotions on FB (23)	(v) Emotion (positive)	4	9	**13**
	(vi) Emotion (mixed)	3	3	**6**
	(vii) Emotion (negative)	3	1	**4**
		(10)	**(13)**	**(23)**
(III) Perceived effect of FB (54)	(viii) Recognition of errors/mistakes	12	1	**13**
	(ix) Realization of anomalies	6	7	**13**
	(x) Chance to reflect on one's writing	3	7	**10**
	(xi) Obtaining an objective viewpoint	2	6	**8**
	(xii) Benefits for the future	4	3	**7**
	(xiii) Achieving succinct modifications	0	3	**3**
		(27)	**(27)**	**(54)**
(IV) Critiques (14)	(xiv) Critiques for FB	6	3	**9**
	(xv) Critiques for class	3	2	**5**
		(9)	**(5)**	**(14)**
Total		**71**	**69**	**140**

Note. The table compares the control group (*n* = 22) and intervention group (*n* = 23). FB, feedback

First, the similarities between the two groups as seen in Table 4.26 are summarized. A common point between the two groups is that they felt the feedback was more comprehensive than the feedback they had received in the past, as shown in code (iv), comparison to FB received in the past. Respondents who made this comment may have received similar feedback for EFL writing instruction in junior or senior high school. Moreover, some students recognized and appreciated the effort of their teacher in giving feedback on so many students' writing several times. These comments are included in positive emotion of code (v) and show no difference in frequency and content between the two groups. Next, regarding code (vii), benefits for the future, in theme III, it is interesting in that both groups had about the same frequency of comments, 4 and 3, respectively, and both groups shared the feeling that they could take advantage of this feedback, but they differed slightly in their strategies. An example comment from the control group states, "The feedback pointed out points that were difficult to notice in my own study, so I thought I could make use of it in my future writing." (自分で学習するうえでは気づくことが難しい点を指摘してもらえたので今後のライティングにも活かせると思えた) A example from the intervention group says, "What I've learned from the feedback given to me, it was important to read my sentences over and over again and correct them if I think something is even slightly wrong, in my future as well." (フィードバックをもらい学んだことは、何度も読み直し少しでもおかしいと思ったら修正して行くことが今後の作文でも大切だと思った) It can be inferred from these comments that the control group is trying to utilize the specific knowledge learned from the feedback, while the student in the intervention group thought highly of and attempted to take advantage of the revision process of repeated reading. Moreover, both groups shared code (vi), mixed emotion. As can be seen from the excerpts in Table 4.25 the students honestly express their joy that they were able to write a satisfactory sentence by referring to the feedback and revising their essays, about which they were not confident in the beginning. There is no difference between the two groups in this sentiment, either in the frequency or in the content.

While there were some similarities between the two groups as described above, there seemed some obvious differences regarding the feedback. First, there was a notable difference in the number of comments related to codes (i) and (ii) of theme I between the groups. In terms of student evaluation of the feedback, 27 comments focused on the clarity of the feedback, 74% of which (20 comments) came from the intervention group. Conversely, 11 (85%) of the 13 comments focused on the detail of the feedback. This may be due to the fact that the schematic feedback helped to facilitate understanding of the entire text at a glance, whereas the conventional feedback in the control group is interspersed throughout the text with individual points of view, giving the impression that the feedback is detailed. In fact, the result is interesting in that few (only two instances) students in the intervention group commented that the feedback was detailed even though the grammatical and lexical errors were also noted to the intervention group. This may suggest that the intervention group was more conscious of the diagram itself in the feedback than individual errors pointed out in terms of language use. This also relates to the emotions of theme II. The control group had a slightly higher frequency of negative emotions than the intervention group, and at the same time had a lower frequency of positive emotions. Some of the control group students seemed perplexed when they had many "mistakes and errors" pointed out, which was observed in the excerpt in the table above as well as the following, "In revisions based on feedback, it was difficult to adjust to the specific weight of the words, and I was anxious about the overall balance and whether I complemented the words properly." (フィードバックに基づく修正では、単語の比重との調整が難しく、全体のバランスを意識してしまい、うまく補足できているのか不安だった) Another example from the control group is: "What I think is not well put into English. On my own, I'm not confident that I can notice mistakes to know where exactly to fix them. I felt a lack of knowledge and English language skills." (思っていることがうまく英語に置き換えられない。自分だけではどこを具体的に直せばいいか間違いに気づける自信ない。知識不足や英語力のなさを感じた) This comment is mainly a reference to language use. This may suggest that themes II and III are interrelated to each other to some extent.

Next, with regard to code (x), chance to reflect on one's writing, there were seven comments for the intervention group and three for the control group. One interesting comment from the intervention group is presented for discussion. The student commented, “As in normal conversation, I tend to ramble on for a long time, and I think I'm talking about A → B → C, when in fact I'm talking about A → A' → A", and I tend to take a step back on the topic, so I need to be careful.”（普通の会話でもそうだが 僕は一言が長いようで、A→B→C と話しているようで A→A‘→A”と話題の足踏みをしがちで注意したい。）It is intriguing to see how he not only analyzed his own writing style, but also assessed his reasoning objectively when he made his opinion statements.

The analysis regarding code (xiii), achieving concise modification, should be discussed keeping in mind that, as a premise, the frequency of occurrence is so small (three in the intervention group) that it cannot be generalized and should be treated with caution. Code (xiii) was found in three cases only in the intervention group, but the remaining instance, not shown in the table, was “I couldn't express what I wanted to say while writing initially and it was worded in a roundabout way, but I think, thanks to the feedback, it became clear and coherent.”（１回目は、書きながら言いたいことがうまく表現できなくて回りくどい言い方になってしまったがフィードバックのおかげでスッキリとまとまったものになった。）This “succinctness” is also reflected in the decreased word counts and sentence units in the writing in the posttest results.

Finally, for theme IV there were more instances extracted for critiques from the control group, especially with respect to feedback, than from the intervention group, but some comments do not necessarily seem to be attributed to different forms of feedback, while others seemed specific to each group. As comments on critiques are important for future instructions, excerpts other than the two examples shown in Table 4.25 are also presented here. The following four comments are from the control group:

> I would have liked to see what kind of sentences the teacher would have written in response to this topic. (改善部分の実際の例があるとより分かりやすかったと思う)

> I would have liked to see what kind of sentences the teacher would have written in response to this topic.（先生だったらこのお題に対してどんな文を書くのか見てみたい）
>
> I would have liked to see a rough estimate of how many points I would have received if I had been given a score.（点数を付けていただいた時にどれぐらいだとこの点数になるのかなと目安の文なども見てみたい）
>
> I would have liked to see the teacher read and evaluate a couple more paragraphs.（もうあと 1，2 個作文を読んで評価してもらいたかった。）

Except for the last comment, the critique was to a request for a model writing sample for the assignment topic. Because this was a research project, the timing of the presentation of the model writing missed the opportunity to present the model writing, but it will be necessary to consider appropriate timing and ways to provide the model writing in the future.

The following three comments are from the intervention group. The students requested further feedback on language use:

> I felt that I could have learned more if you had shown me some other ways of writing and expressing myself when correcting the English sentences.（英文の修正のところで何個か別の書き方や表現方法を示してもらえるとさらに勉強になったかなと感じた).
>
> I have forgotten a lot of grammar and word usage, so if there are any mistakes in the correction, I would like to know them too.（文法や語法を結構忘れているため添削でミスがあったらそこも教えていただきたい）
>
> I had difficulties using the tips and suggestions other than those given by the teacher. (先生が出された案以外を使いづらかった)

Although feedback on language use was included, students who were anxious or interested in language use may have felt that it was insufficient. The last comment from the intervention group showed next suggests that any teacher feedback could possibly be one-sided and an intrusive. It seemed to show the difficulty of teacher writing feedback. No matter how carefully one makes a recommendation, recommending something means that something else cannot be recommended.

Having discussed some of the important points from the student comments and the aggregate results of the thematic analysis above, it can be said that the study participants generally had some positive emotions in response to the both text-based teacher feedback to the control group and the schematized feedback with an annotation diagram to the intervention group. However, the comments from the control group were more abstract:

> It was fresh.（新鮮だった）
>
> I was happy (to receive feedback) even though I don't know exactly why myself.（なんだかわらないがうれしかった）
>
> I was motivated.（モチベーションが上がった）.

On the other hand, some comments from the intervention group clearly showed more interest in the schematized feedback:

> It was interesting that each person's diagram was different.（一人一人図式化した形が異なっていて興味深かった）
>
> It was interesting that I could see the object assembled in my brain as a form. (図式化 されることで自分の脳内で組み立てられたモノが 形 として見ることができたのがおもしろかった）

In addition, there were comments indicating a sense of accomplishment such as:

> It was nice to be able to correct unsatisfactory parts with feedback successfully eventually.（納得がいかなかった部分を最終的にフィードバックで直せるのは気持ちいい）
>
> I was able to create a satisfactory text.（満足のいく文章が作れた）.

In this section, the two groups that have received different forms of teacher feedback, the control group and the intervention group, have been analyzed from two perspectives: the changes they made in the rewriting task and their impressions of the feedback itself. There were similarities and differences between the groups, but the comments related to theme IV (critique) seem to symbolize the two groups. In other words, both groups benefited from the feedback and had positive impressions, but the control group's comments were abstract and holistic, whereas the intervention group's comments were individual and specific. It seems clear that the new form of schematized feedback brought new stimuli to some students, and that there were differences in the way they made corrections and in the way they felt.

4.2.4.3 Creating a Tentative Storyline Derived From the Summary of Themes. As a final product of the thematic analysis, Creswell and Creswell Báez (2021) suggest creating a storyline illustration based on the extracted themes. Following their approach, a tentative storyline was created by inferring how the extracted themes by the thematic analysis are related to each other. The intention is that an illustration would provide a better understanding of the effect of the teacher feedback on students' rewriting process. Figure 4.15 explains a series of students' behaviors in the process of rewriting based on the four themes.

As described in theme I, when a student receives feedback, he/she would first evaluate the format and the content of the feedback to examine how clear (code i) it is to understand, how detailed and thorough (code ii) to be helpful for revision of the writing, by comparing it with the one they had received in the past (code iv) subconsciously. In other words, the students who receive teacher feedback examine the feedback. At this point, they would decide whether they agree with or understand the feedback they have received, and after determining what and how to fix it, they begin the revision process.

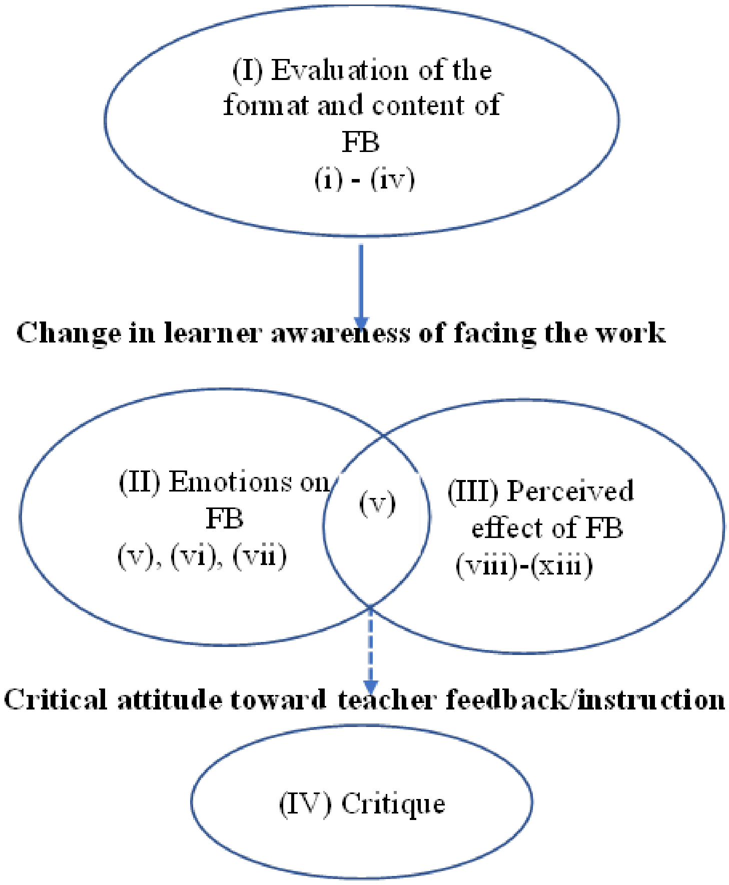
In doing so, they seem to experience some positive emotion (theme II, code v) of respect for their work, joy at having their English writing understood, or confusion (code vi) at their own lack of English writing ability and satisfaction at having made a successful correction, and sometimes feel anxious about their work or overwhelmed by its difficulty (code vii).

At the same time, the students seem to have experienced a sense of growth and accomplishment by recognizing problems in their own work through feedback that they could not have recognized on their own without it (theme III, codes viii and ix). With teacher feedback, they came to look at their own writing work objectively (codes x, xi) by being evaluated by and receiving comments from someone who they can trust in terms of English language learning. They occasionally discovered what they could apply to future writing (code xii). These findings suggest that, through the revision process, the students had the opportunity to face their work with a different attitude than before (code xiii).

Eventually, some students, based on their experience, might feel that they have further demands on the faculty, that is, they are critically confronted with (or are able to verbalize) teacher feedback, which should be the most favorable consequence for the instructor (theme IV).

Figure 4.15

A Storyline Illustrating the Relationship Between the Themes


(I) Evaluation of the format and content of FB
(i) - (iv)
Change in learner awareness of facing the work
(II) Emotions on FB
(v), (vi), (vii)
(v)
(III) Perceived effect of FB
(viii)-(xiii)
Critical attitude toward teacher feedback/instruction
(IV) Critique

Note. Based on Creswell & Creswell Báez (2021). FB, feedback

While taking these steps, the students receive feedback, which is the product of formative assessment: They are to understand, evaluate, and judge the feedback and actively take the initiative to make appropriate revisions. In this process, the learner's awareness of their own work would change, and some learners might even come to think critically about the feedback and instruction. The revision assignment based on feedback in formative assessment seems to be a catalyst that encourages students to act independently. Furthermore, while both the control and intervention groups seemed to have taken essentially the same steps in this process, there were certain differences in how they experienced and felt about the process, as well as in how they revised their writing.

Chapter 5

Discussion

5.1 Introduction

The purpose of the present study was to examine the effectiveness of giving schematized teacher feedback on stand-alone paragraph writing to undergraduate students in Japan. This chapter discusses the three RQs by integrating the quantitative and qualitative results from the mixed methods approach. Fundamentally, a mixed methods approach is highly appropriate for communicative language assessment (Moeller, 2016). There are quite a few variables in the classroom assessment, where instruction and testing are often separable; hence, a mixed methods approach is suitable to interpret students' performance.

While a mixed methods approach was adopted as the overall research design (Creswell, 2015; Kakai, 2015; Miller & Bustamante, 2016), the RQs and procedure were established based on the Assessment Use Argument (AUA) framework (Bachman & Palmer, 2010; Bachman & Damböck, 2017). In the AUA, the assessment study proceeds sequentially based on the assumption of the previous claim by examining each intended claim.

The RQs in this study were addressed by following this assessment procedure. RQ 1, which evaluated the consistency and appropriateness of the rating scale (Claim 4 in the AUA) was generally confirmed so that the subsequent series of assessment could be implemented. For RQ 2, the results of several studies showed that the interpretation of the scores provided raters, annotators, and students with information about the ability to be assessed. In addition, the annotated tree diagrams, which are another form of providing assessment, indicated the meaningfulness of the interpretation (Claim 3 in the AUA). Last but not least, the studies aimed to address RQ 3 provided quantitative and qualitative data about the effectiveness of schematized feedback (Claim 1 in the AUA). The following sections summarize each RQ one at a time.

5.2 RQ 1: The Reliability of the Rating Scale and Annotation Scheme (Claim 4 in the AUA)

The first RQ concerned validation of the assumptions underlying the subsequent analyses, which address Claim 4 in the AUA framework: the consistency of the rating scales and annotation schemes used in this study. As a quantitative analysis, a Many-facet Rasch Measurement (MFRM) analysis was employed to examine the quality of the modified analytic rating scale of the ESL Composition Profile (Jacobs et al., 1981) on writing skill criteria—content, organization, language, and vocabulary—over three writing occasions on two tasks. In addition, a transfer task was employed to examine the retention of writing skill(s) for the last occasion.

Among Linacre's (1999, 2004) six guidelines for checking the quality of the rating scale, there were two concerns in the rating scale employed in the present study: the size of threshold increase (≥ 1.4 and < 5.0 logit) was not always satisfied, and the number of responses at level 1 was < 10 for some criteria. Although one of the remedies for these violations is to collapse categories, this approach was not employed in the present study. The rationale behind the decision was that the raters virtually rated the students' writing samples at six levels for scoring. As described in the Section 3.2.2.1.2, the original ESL CP contains a 100-point rating scale and a four-level ordinal scale of evaluation criteria. However, the author's follow-up discussions with the raters revealed their actual rating behaviors recognized six levels. It should be important for the validity and reliability of the rating scale to be as close as possible to the reality of the rater behaviors, especially in the classroom assessment. Additionally, Eckes (2015, p. 121) proposes that it is not always a good idea to take an easy action of collapsing categories, in particular with small samples of raters and/or examinee performances. Therefore, the 6-point rating scale used in this study satisfies a certain degree of validity for use in a low-stakes assessment context in the classroom. However, as a subject for a future study, it would be desirable to revisit the rating scale using an even larger sample size or with another population.

In terms of inter-annotator agreement, Putra et al. (2021), the creators of the TIARA annotation tool, state that their “argument annotation scheme is demonstrably stable, achieving good inter-annotator agreement and near-perfect intra-annotator agreement” (p. 1). Still, in order for annotators to understand each other’s coding scheme and to achieve a certain level of agreement, it is true that quite careful preparation is required before the actual coding. This endeavor includes understanding the concept, performing pilot studies in advance, and establishing a communication system to resolve any problems. As with any empirical study, the above is necessary for assessment studies using annotation tools.

Regarding the consistency of the annotation tool used in the present study, which is presented in Section 3.2.5, two basic coding schemes for inter-annotator agreement and two other coding schemes specifically related to the issue of this study were checked. Thus, a total of four inter-annotator agreement levels were examined. The basic ones are the source-target agreement, 86%, and the relation labeling agreement, 81%. The study-specific ones are the location of the anomalous units in discourse, 82%, and the labeling of the anomaly type, 71%. The degree of agreement for decisions on the type of anomalies was lower than that for decisions on location since multiple interpretations were possible. Possibility of multiple interpretations means that there are multiple ways of modifying anomalies. Since the feedback to the students only marked the location and of the anomaly as a relation label, but not the anomaly type, it was up to the student to decide how to modify it. This could be the pedagogically important aspect of schematized feedback.

As noted above, there were some concerns that needed to be approached with caution when making judgments. Overall, however, both the writing rating scale and the annotation coding scheme were found to be reliable. Accordingly, Claim 4 of the AUA was supported by this study.

5.3 RQ 2: The Meaningfulness of Interpretation of Students' Performance (Claim 3 in the AUA)

The second RQ, which addressed Claim 3 of the AUA, examined whether the interpretation of the student performance evaluated on the rating scale as well as the annotated tree diagram as an assessment record were meaningful in the study. The former required quantitative analysis and the latter required qualitative analysis. For the quantitative analysis, two perspectives were incorporated: the total score across all the criteria presented on a 100-point scale, and the score of each criterion based on the 6-point scale that had been confirmed reliable (based on RQ 1). Parallel coordinate plots were also utilized.

Regarding the total score across all the criteria, in addition to the presentation of descriptive statistics, boxplots and parallel coordinate plots were examined to visually inspect the overall results. In particular, parallel coordinate plots, also known as profile plots, were informative in that they showed the diversity of students in the classroom, which is often overlooked by averages alone.

First, the MFRM task difficulty results were the most straightforward indication of the meaningfulness of the interpretation of the scores. In terms of difficulty, the occasion-by-task combinations were ordered from the most difficult to the easiest: Pre B > Pre A > Transfer C > Post B > Post A (easy). Of note, the second draft (denoted as Post), which was a revised draft of the first draft (denoted as Pre) with teacher feedback, increased the score, and the Transfer C task was in between; Topic B was more difficult than Topic A. Considering that Topic A was taken from Eiken Grade 2, and Topic B from Grade Pre-1, this finding makes sense because the Grade Pre-1 level is originally set at a higher level of difficulty than the Grade 2 level. In addition, it is not surprising that the transition task in the new prompt is more difficult than the revised task, which is the second draft. Therefore, the scoring results for the combination of these two factors were meaningful.

The results in terms of total points are also consistent with the FACET results above in that writing scores clearly increased with teacher feedback regardless of the task. Moreover, Topic B, which was more difficult, exhibited a consistent increase in scores due to the lower Pre B task scores, and the drop was smaller than that on Topic A, even in the Transfer C task. Regarding the parallel coordinate plots, at the individual level, some students' scores shifted differently from the overall group. The issue of individual differences is important in the classroom, but it is not always easy to identify the factors that contribute to these differences. Individual differences may be due to differences in feedback types, writing skills criteria, more personal factors, or a combination of these factors, and should be considered in light of the results of the qualitative surveys. Regarding the interpretation of the writing skills criteria, while the retention effect was limited for the language and vocabulary criteria, there was a certain degree of retention for the content and organization criteria.

Lastly, regarding the meaningfulness of the interpretation of the annotation scheme, because there is no standard examination method, the original method was developed to check the correlation between the number of coherence anomalous units in a writing sample as identified by the annotator using an annotated diagram and the organization score. Because the frequency of anomalies contained in a writing is not the only factor that determines the quality of coherence, it cannot be used as an absolute indicator. However, generally speaking, writing that contains more anomalies should have a lower organization or content score. Therefore, in this approach, the relevance or meaningfulness of the AUA annotation scheme could be examined as a supplemental indicator. The correlations between the total anomaly frequencies and the content and organization scores were negative and significant, -.497 and -.511, respectively, suggesting that the annotation scheme in this study presented the intended interpretation to a certain extent. As a result, an annotated tree-shaped diagram provided a good representation of the students' language ability in a very easy-to-understand way. The students' responses to the questionnaire on schematized feedback confirmed this view.

5.4 RQ 3: The Effectiveness of Schematized Teacher Feedback (Claim 1 of the AUA)

The effectiveness of schematized teacher feedback created with the annotation tool in the intervention group was investigated by comparison with the control group from the following three perspectives: revision time on task, overall writing performance and organization of their writing across occasions, and students' rewriting behaviors and perceptions. The students' rewriting behavior is related to the decision-making process of Claim 2 in the AUA, that is, what corrective actions students take after receiving feedback. However, the questionnaire did not allow collecting data to adequately support Claim 2, so the analysis was conducted from the perspective of the benefits of Claim 1 of the AUA.

5.4.1 RQ 3(1): Performance on Writing Skills Criteria Across the Occasions (Claim 1 of the AUA)

The effect of schematized teacher feedback is discussed both quantitatively and qualitatively to evaluate the overall writing performance and organization specifically. For quantitative analysis, a mixed-between-within MANOVA was conducted. This analysis is applied in mixed methods research when it includes a between-group independent variable (two different groups based on feedback type: conventional versus graphic) and a within-groups independent variable (three writing occasions in the repeated-measures design). Because a MANOVA considers correlations among the dependent variables, it is appropriate in the case of this study, where the four dependent variables of writing ability are presumed to be related to each other. There were no significant differences between the intervention group, which was given graphical feedback, and the control group, which was given conventional text-based feedback. However, in some cases, there were differences in the transfer task or among the criteria. Interestingly, while differences between the two groups in the organization score were expected, there was actually a larger difference in the language and vocabulary scores. While it is difficult to pinpoint the cause of this phenomenon, one possible explanation is that the annotation diagrams used as feedback were separated by text units, making it easier to focus on a single sentence and to find errors at the sentence level. It is also possible that more attention was paid to each individual word. This may be the same reason that when an annotator "reads" an initial draft, the tree diagram focuses more on a single sentence and makes it easier to read the sentence critically. This may be supported by the fact that the intervention group responded more frequently and in more detail to questions in the qualitative study questionnaire in which they were asked to recall and list the areas of correction.

5.4.2 RQ 3(2): Revision Time on Task (Claim 1 of the AUA)

Secondly, the time required to revise and rewrite the initial writing with reference to the given feedback was measured. This study is based on the theoretical background of previous research on the effect of graphic display in learning. As described in Section 2.5.1.1, Winn et al. (1991) showed that students spent less time working on problem solving in the intervention group presented with graphics than in the control group given textual explanations. In the present study, the mean time spent on rewriting work was about 7 minutes shorter in the intervention group (26.7 minutes) than in the control group (19.3 minutes) for Topic A and about 7 minutes shorter for Topic B (control: 26.5 minutes; intervention: 19.3 minutes). However, the *SD* of the intervention group was larger than that of the control group for both tasks. Furthermore, for Topic B, the students who took the longest spent about the same time between groups (control: 43.0 minutes; intervention: 44.0 minutes). It is not easy to generalize this result because it may also be attributed to each student's cautiousness. As Winn et al. (1991) stated as a limitation, in order for graphic presentation to be effective, students must be familiar with how to view and interpret diagrams, and some students may not prefer explanations in diagrams. To summarize, while there were differences at the individual level, it can be concluded that the presentation of graphics as feedback helped the intervention group to produce the same number of words of writing in a shorter amount of time. That is, the intervention group performed the rewriting work more efficiently than the control group. The visual argument hypothesis that presupposes "graphical representations are effective because, owing to their visuospatial properties, their processing requires fewer cognitive transformations than does text processing and does not exceed the limitations of working memory" (Vekiri, 2002, p. 281) led to one of the RQs in the present study related to time on task.

5.4.3 RQ 3(3): Ideational and Rhetorical Coherence in the Transfer Task (Claim 1 of the AUA)

In addition to statistical analysis with scores obtained from the rating scale, a series of coherence-specific analyses through annotated diagrams as another indicator of the writing performance revealed several effects presumably specific to the intervention group provided with schematized feedback. The writing quality in terms of coherence was evaluated by performing coherence anomaly analysis and by evaluating the shape of the annotated diagrams. There were four indices to examine the writing: the total frequency of anomalies in writing, the anomaly categories, the types of anomalies hindering coherence contained in writing, and the shape of the tree diagrams.

There were 59 anomalies in the control group (*n* = 22) and 41 in the intervention group (*n* = 23). While there was a certain degree of difference between the control and intervention groups in the total frequency of anomalies, the chi-square goodness-of-fit test showed no significant difference. Regarding the location of the anomaly, there was no significant difference between the control and intervention groups overall, but there were two distinct differences. First, the intervention group had seven zero-anomaly writing samples, while the control group had four. Second, the intervention group had zero counterargument anomalies, while the control group had four.

In terms of the 11 anomaly categories, both the control and intervention groups presented the most in Category 1, poor reasoning, logic, and topic development, namely

24 in the control group and 10 in the intervention group. There was also a high number of Category 6 anomalies, irrelevant support, 15 in the control group and 7 in the intervention group. However, the relationship between the anomaly category and score could not be generalized due to the small sample size.

Lastly, the shapes of the tree diagrams generated with the TIARA annotation tool were classified into four types according to the general structural features: well-balanced, horizontally wide, unbalanced, and vertically long. The first represents a well-formed diagram, while the other three are poorly formed diagrams. The percentage of the three types of poorly formed tree diagrams was higher in the control group (50%) than in the intervention group (30%). This finding suggests that the intervention group produced writing represented by well-balanced diagrams that led to higher organization scores than the control group. A paragraph with a balanced form indicates that the components of the paragraph have been met, such as a good balance in the number of major and minor supporting sentences, as well as an alignment of counterarguments and rebuttals.

5.4.4 RQ 3(4): Students' Rewriting Behaviors and Perceptions (Claim 1 in the AUA)

First, in the questionnaire, students were asked to list, for each writing skills criterion, as many revisions as they could think of when completing the second draft in response to the feedback. This leads to the ability to respond to revisions and to come up with options for revisions to make a better draft by referring to the teacher feedback. This question addressed Claim 1 of the AUA.

The students reported 173 revisions, 78 from the control group and 95 from the intervention group. The main cause of this difference between the two groups was content and organization (control group: 21 responses; intervention group: 40 responses); the frequencies for the other criteria were about the same for each group. Next, thematic analysis was used to extract more detailed codes from the responses in content and organization. These codes were merged into four themes: (a) addition/supplementation, (b) revision/alteration, (c) rearrangement of sentences, and (d) deletion. The most frequent difference between the control and intervention groups was in (b), revision/alteration (control group: 5 responses; intervention group: 14 responses). More interesting, however, is the fact that the intervention group's responses corresponded to all eight codes, whereas the control group reported only five, indicating that they reported fewer types of revisions. These results may indicate that the intervention group was more responsive to the feedback.

Finally, thematic analysis was employed to evaluate the students' responses to the open-ended questions about how they felt about the feedback. There were 45 respondents, 22 from the control group and 23 from the intervention group. Each comment was segmented into units for analyses. A total of 140 units were identified, 71 from the control group and 69 from the intervention group. The resulting units were coded in an exploratory manner. As a result, a total of 15 codes and four themes based on those codes were identified. There were both similarities and differences between the results of the two groups. One of the common points between the two groups were that they felt the feedback was more comprehensive than the feedback they had received in the past, as shown by code iv (comparison to feedback received in the past). Respondents who made this comment may have received similar feedback for EFL writing instruction in junior or senior high school. Moreover, some students recognized and appreciated the effort of their teacher to give feedback several times on so many students' writing samples. There seemed to be no difference between the groups in the sentiment, either in the frequency or in the content. One point where the two groups differed was in their evaluation of the feedback. The control group evaluated the feedback as "detailed," whereas the intervention group evaluated it as "clear" or "easy to understand." This difference may be due to the fact that the schematic feedback helped to facilitate understanding of the entire text at a glance, whereas the conventional feedback in the control group was interspersed throughout the text, giving the impression that the feedback was detailed.

There is one comment from an intervention group student that is very symbolic, which is the final example. The comment was categorized into code x (chance to reflect on one’s writing). He commented, “As in normal conversation, I tend to ramble on for a long time, and I think I'm talking about A → B → C, when in fact I’m talking about A → A' → A", and I tend to take a step back on the topic, so I need to be careful.”（普通の会話でもそうだが　僕は一言が長いようで、A→B→C と話しているようで A→A’→A”と話題の足踏みをしがちで注意したい。）It is intriguing to see how he not only analyzed his own writing style, but also assessed his reasoning objectively when he made his opinion statements.

5.5 Limitation of the Study

Empirical research conducted in the classroom always suffers from an educational ethical perspective. It is not always easy to apply strict conditions to the experimental and control groups, which is often an important condition for assessment research. The greatest possible care should be taken to ensure that neither group is disadvantaged. In this study, the independent variable was the form of teacher feedback (a diagram or conventional text). When explaining the Toulmin model of argument (Toulmin, 2003) as a part of the classroom instructions, which was referred to as the argument model in the prior study, a graphical explanation was used in both groups. This was inevitable due to the nature of the Toulmin model, but the form of this diagram is naturally related to the annotated diagram in the task writing. It is not clear whether the presentation of this diagram had any subconscious effect on the control group.

Another issue is sample size. The EFL writing classes in Japanese universities are often small because of the time it takes to review assignments. Although it is possible to compensate for this issue by continuously teaching similar classes for several years and conducting similar empirical studies, it is difficult to ensure an experiment that comprises a large, homogeneous population and sufficient statistical power. This limitation always accompanies empirical studies in the classroom.

Finally, the fact that the authors themselves were classroom teachers who were familiar with the two classes in this study may have been a factor in the qualitative changes in student performance. This is a point where generalization of results should be cautious.

5.6 Pedagogical Implications

It is highly recommended to reaffirm the value of graphic displays in learning and to utilize them in some form in the classroom. As noted in Section 5.4, the creation of diagrams with annotation tools leads to significant learning and awareness about coherence for the learner and the annotator. If one does not feel comfortable using annotation tools, it is possible to create diagrams by hand for stand-alone paragraph writing. Alternatively, just presenting and explaining a model tree diagram would be effective. Furthermore, as evidenced by the students' answers to the questionnaire, some students mentioned the effect of reading their own writing, which was segmented into one sentence at a time in the box. This suggests that simply listing the segmented sentences one by one may be effective to focus attention on a single sentence.

Finally, regarding the teacher's effort in creating the feedback, since this study used two tasks with different levels of difficulty, the feedback was created twice. However, the students' comments indicated that the graphical feedback left a strong impression, and therefore it is considered effective even when implemented once per semester.

5.7 Directions for Further Research

There are two future directions for the present study. One is a study of the effectiveness of instructors' use of annotation tools to teach ideational and rhetorical coherence in EFL writing. In this study, the annotation was validated by the author and another researcher, who also used this annotation tool within a writing class, so there is a possibility for future research on the instructor that the diagramming work can bring.

In order to make the study more objective, it would have to involve even more researchers.

Another possibility is to have students use the annotation tool themselves. As described earlier, the annotation tool involves a series of tasks, starting with segmenting the text into single sentences, tagging text units, and connecting source text to target text, which should provide an excellent experience for thinking about coherence. It would be a unique experience to think about coherence. The annotation tool used in this study is freely available on the Internet, lightweight, and extremely easy to use. However, it must be noted here that, as pointed out in the results of this study, there were some students who were not suited to the graphical understanding due to unfamiliarity or natural aversion to diagrams, and similarly, it is anticipated that there will be some students who are not suited to working with the software due to poor use of computer software. It would be necessary to consider an alternative plan for them.

5.8 Summary of the Discussion

While the quantitative results showed a few significant differences between the two groups, the qualitative study showed some more positive effect in the intervention group given schematized feedback in terms of their perception of the revision process, their perception of the feedback, and the frequency of the coherence anomaly and the shape of the annotation diagram. It is not clear whether the lack of significance between the control group given conventional text-based feedback and the intervention group given schematized feedback in terms of quantitative analysis was due to the actual lack of significant differences in feedback effects or other issues, such as the raters, the rating scale used in this study, or the interaction of all of these factors. Moreover, a small sample size decreases the statistical power, so it may be difficult to statistically confirm significant effects in classroom assessments.

Chapter 6

Conclusion

The present mixed methods study addressed the issues regarding the teaching and learning of ideational and rhetorical coherence in English as Foreign Language (EFL) paragraph writing in terms of formative classroom-based assessment of university students at a basic level in Japan. The study was conducted to investigate the efficacy of schematized feedback with a tree-diagram generated by an annotation tool. The results showed that there was no significant statistical difference between the two groups, except for some results in the transfer task. On the other hand, the results of qualitative study showed some positive effect with the intervention group provided with the schematized feedback such as the reduction of time on task during revisions, the well-balanced shape of annotated diagrams in transfer task, which indicates a high-quality organization of the passage. Furthermore, the results of a questionnaire showed that the intervention group students tended to be more sensitive and attentive to writing coherence than the control group.

Despite some limitations associated with empirical research conducted in a classroom setting, such as small sample size, difficulties in controlling the educational environment among control groups from an ethical standpoint, this study provided several pedagogical implications. One of them is the effectiveness of graphic displays for EFL writing activities in teaching and learning. Furthermore, since the annotation process itself, such as text segmentation, tagging, and connecting text units, may provide an opportunity for L2 writers to improve the understanding of coherence, it would be worthwhile to conduct future research on further utilization of annotation tool by instructors and on students' own attempts at engaging in annotation work.

The value of graphical displays in learning and teaching in EFL writing should be discussed here. As was described in Chapter 2, Vekiri's (2002) conducted a comprehensive review of research on the value of graphical displays in learning and presents a number of studies that investigate the potential power of graphic displays in learning. Indeed, the author of the present study prefers to glance at figures and tables rather than read a long explanatory text. The motivation for this study was related to such personal preferences. In fact, many people have experienced that it is faster and easier to understand a graph, diagram, or table than to read a long explanation in writing. The reason why it is easier is probably because the creators of figures and tables have spent a certain amount of time beforehand in the process of understanding, digesting, and effectively presenting the content to be conveyed; in other words, they have taken care to prevent their work so that it is easily understandable. The same is true for the presentation of the annotated diagram in this study as feedback. Although the quantitative investigation of this study did not always find significant differences, the qualitative investigation showed that the corrective behavior and perception in the intervention group that received the schematic feedback was clearly different from the control group, with some students' comments such as "the feedback is easy to understand." This view is probably due in part to the annotators' prior experience and efforts with providing graphical feedback. This may also be attributed to the visual argument hypothesis that presupposes the viewers' information processing requires fewer cognitive transformations than does text processing (Vekiri, 2002). Furthermore, when the transfer task writing samples of the intervention group were converted into an annotated diagram, there were more balanced, model-like tree diagrams in the intervention group than in the control group. This may suggest that the students possibly have had an ideal diagram in their mind when composing their writing. In fact, during the planning phase of working on the transfer task, some students drew their own diagram on paper to work prior to the actual writing. The image of the diagram as the ideal form of the paragraph structure proposed in the schematic feedback seems to remain as a visual image in the learners' minds. Regarding the procedure of annotation, the creation of feedback using annotation tools is an experience of learning and discovery not only for the learners, but also for the annotators themselves, and the significance of classroom instructors using this approach is not small. Above all, the following comment from a student in response to the feedback reflects its value well: “It was interesting that I could see the object assembled in my brain as a form.” It also resulted in a change in learners' writing behavior, as some students in the intervention group were observed to use the learning strategy of drawing their own tree diagrams during the preparation phase of the transfer task.

In closing, the research design of this study is discussed in terms of coherence. This whole study followed the AUA framework and mixed methods approach. The AUA framework is oriented toward assessment use justification, which examines to what extent and how the intended consequences are beneficial to the participants. If the issue of this study is ideational and rhetorical coherence in writing, then this assessment framework is an assessment of coherence that links the intended assessment record, intended interpretation, intended decisions, and intended consequences in a bead-like chain. In writing annotation, the link from the source text to the target text is the basis of coherence, but just as a passage is coherent only when it is in harmony with the whole, in research, the relationship between each study and the positioning of the research as a whole should always be considered. In research, it is essential to always be aware of the relationship between each study and the positioning of the research as a whole. Through the process writing assessment as one of the classroom activities in this study, it is hoped that the students have learned the importance and essence of coherence in any project. As discussed above, quantitative results alone may not have supported the effect of schematized feedback, but qualitative analysis showed an overwhelming advantage to the intervention group in terms of the soundness of the anomaly analysis and the shape of the tree diagram, or the questionnaire responses. Moeller (2016) notes, “The numerous variables and complexity in assessing authentic task-based communication at the classroom level, one research method cannot fully capture the complexity of language skills” (p. 8). Here lies the value of a mixed methods approach in classroom-based assessment.

References

Adamson, G., & Bunting, B. (2005). Some statistical and graphical strategies for exploring the effect of interventions in heal research. In J. Miles & P. Gilbert (Eds.), *A handbook of research methods for clinical and health psychology* (pp. 279-294). Oxford University Press.

Ahmadi, A. & Parhizgar, S. (2017). Coherence errors in Iranian EFL learners' writing: *A Rhetorical Structure Theory Approach. Journal of Language Horizons, 1*, 9-37. https://doi.org/10.22051/lghor.2017.8588.1011

Alderson, C. (2005). Diagnosing foreign language proficiency: The interface between learning and assessment. London: Continuum.

Allwood, J. (1981). On the distinction between semantics and pragmatics. In W. Klein & W. Levelt (Eds.), *Crossing the boundaries in linguistics* (pp. 177-189). Reidel.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association. Retrieved from https://www.apa.org/science/programs/testing/standards

Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing, 9*(3), 227-257. https://doi.org/10.1016/S1060-3743(00)00027-8

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing 17*(1), 1-42. http://dx.doi.org/10.1191/026553200675041464

Bachman, L. F. & Damböck, B. (2017). *Language Assessment for Classroom Teachers.* Oxford University Press.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238–257. https://doi.org/10.1177/026553229501200206

Bachman, L. F. & Palmer, A. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice.* Oxford University Press.

Bamberg, B. (1984). Assessing Coherence: A Reanalysis of Essays written for the National Assessment of Education progress, 1969-1979. *Research in the Teaching of English, 18* (3)*,* 305-319. National Council of Teachers of English. https://www.jstor.org/stable/40171021

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2)*,* 86-107. https://doi.org/10.1016/j.asw.2007.07.001

Belmonte, A. & McCabe, M. (1998). Theme-Rheme patterns in L2 writing. *Didáctica (Lengua y Literatura), 10*, 13-31. Retrieved from https://www.researchgate.net/publication/39279565_Theme-Rheme_patterns_in_L2_writing

Benesse (2021). GTEC tests. Retrieved from https://www.benesse.co.jp/gtec/en/

Bertin, J. (1983). *Semiology of Graphics.* The University of Wisconsin Press.

Biber, D., Nekrasova, T., & Horn, B. (2011). The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis. *TOEFL iB^{TM} Research Report, TOEFL iBT-14.* https://doi.org/10.1002/j.2333-8504.2011.tb02241.x

Bitchener, J. & Storch, N. (2016). *Written Corrective Feedback for L2 Development*. Multilingual Matters. https://doi.org/10.21832/9781783095056

Black, P. & Wiliam, D. (1998). Assessment and Classroom Learning, Assessment in Education: Principles. *Policy & Practice, 5*(1), 7-74. DOI: 10.1080/0969595980050102

Bond, T.G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Lawrence Erlbaum Associates.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. http://dx.doi.org/10.1191/1478088706qp063oa

Brentani, E. &Golia, S. (2007). Unidimensionality in the Rasch Model: How to Detect and Interpret. *STATISTICA, 67*(3), 253-261. doi: 10.6092/issn.1973-2201/3508

Candlin, C., Plum, G., Spinks, S., & National Centre for English Language Teaching and Research (1998). *Researching academic literacies.* Macquarie University.

Cerniglia, C. S., Medsker, K. L., & Connor, U. (1990). Improving coherence by using computer-assisted instruction. In Ulla C., & Ann. M. J (Eds.), *Coherence in Writing: Research and pedagogical perspectives.*(pp. 227-241). Teachers of English to Speakers of Other Language, Inc.

Chandler, J. (2004). A response to Truscott. *Journal of Second Language Writing, 13*(4)*,* 345-348.

Chapelle, C. A., Enright, M. A., & Jamieson, J. M. (2008). Test score interpretation and use. In Chapelle, C. A., Enright, M. A., & Jamieson, J. M. (Eds.), *Building a validity argument for the test of English as a foreign language* (pp.1-26). Routledge.

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. Kunnan (Ed.), *The companion to language assessment* (pp.1-17) John Wiley & Sons, Inc. DOI: 10.1002/9781118411360.wbcla110

Cheng, D., & Steffensen, M.S. (1996). Metadiscourse: A technique for improving student writing. *Research in the Teaching of English, 30*(2), 149-181.

Chou, Y. T., & Wang, W. C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement, 70*, 717-731.

Cohen, A.D. (1994). *Assessing language ability in the classroom.* Heinle & Heinle.

Colby-Kelly, C., & Turner, C. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *Canadian Modern Language Review,64*(1), 9-37. https://doi.org/10.3138/cmlr.64.1.009

Connor, U. (1996). *Contrastive Rhetoric: Cross-cultural aspects of Second-language writing.* Cambridge University Press.

Connor, U. & Farmer, M. (1990). The teaching of topical structures analysis as a revision strategy for ESL writers, In B. Kroll (Ed.). *Second language writing: Research insights for the classroom,* (pp. 126-139). Cambridge University Press.

Corder, S. P. (1967). The Significance of Learners' Errors. *International Review of Applied Linguistics in Language Teaching, 5*, 161-170. http://dx.doi.org/10.1515/iral.1967.5.1-4.161

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *International Journal of Speech ,Language and the Law, 1*(1), 27-43. https://doi.org/10.1558/ijsll.v1i1.27

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge University Press. https://rm.coe.int/1680459f97

Council of Europe (2018).*Common European Framework of Reference for Languages: Learning, Teaching, Assessment Companion Volume with Nes Descriptors.* https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Creswell, J.W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson.

Creswell, J.W. (2015). *A Concise Introduction to Mixed Methods Research.* Sage Publication.

Creswell, J. W.,& Creswell Báez, J. (2021). *30 Essential Skills for the Qualitative Researcher* (4th ed.). SAGE Publications, Inc.

Crismore, A., Markkanen, R., & Steffensen, M. S. (1993). Metadiscourse in Persuasive Writing: A Study of Texts Written by American and Finnish University Students. *Written Communication, 10*(1), 39–71. https://doi.org/10.1177/0741088393010001002

Crusius, T.W., & & Channel, C.E. (2004). *Daigaku de manabu giron no gihou* [Argumentation techniques learned in college]. Trans. Sugino, T., Nakanishi,C., & Kono, T.『大学で学ぶ議論の技法』. Keio University Press.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31–51. https://doi.org/10.1177/026553229000700104

Cumming, A., Kantor, R., Powers, D.E. (2001). Scoring TOEFL Essays and TOEFL 2000 Prototype Writing Tasks: An Investigation Into Raters' Decision Making and Development of a Preliminary Analytic Framework, *TOEFL Monograph Series,* MS-22, https://www.ets.org/Media/Research/pdf/RM-01-04.pdf

Cumming, A., Kantor, R., Powers, D.E. (2002). Decision Making while Rating ESL/EFL Writing Tasks: A Descriptive Framework, *The Modern Language Journal, 86 (1), 67-96.* https://doi.org/10.1111/1540-4781.00137

Daneš, F. (1974). Functional sentence perspective and the organization of the text. In Danes, F. (Ed.) *Papers on Functional Sentence Perspective* (pp. 106-128). Prague: Academica. https://doi.org/10.1515/9783111676524

De Kuthy, K., Reiter, N., & Riester, A. (2018). QUD based annotation of discourse structure and information structure: Tool and evaluation. In *Proceedings of the*

International Conference on Language Resources and Evaluation (LREC), 1932–1938. https://paperswithcode.com/paper/qud-based-annotation-of-discourse-structure

Denzin, N.K., & Lincoln, Y.S. (1994). *Handbook of qualitative research*. Sage publications.

Doe, C. (2014). Diagnostic English Language Needs Assessment (DELNA), *Language Testing 31*(4), 537-543. DOI: 10.1177/0265532214538225

Doe, C. (2015). Student Interpretations of Diagnostic Feedback. *Language Assessment Quarterly, 12*, 110-135. https://doi.org/10.1080/15434303.2014.1002925

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments.* Peter Lang Edition.

EIKEN. (2016). Writing Test Scoring Guidelines and Cautions (Grade 1, Pre-1, and Grade 2). Retrieved from https://www.eiken.or.jp/eiken/exam/2016scoring_w_info.html

EIKEN. (2018). Eiken Writing Score and CEFR Level Mapping. Retrieved from https://www.eiken.or.jp/eiken/group/result/pdf/eiken-score-cefr.pdf

EIKEN. (2021). Situation of taking an examination. Retrieved from https://www.eiken.or.jp/eiken/merit/situation/

Elder, C. (2003). The DELNA initiative at the University of Auckland. *TESOLANZ Newsletter, 12*(1), 15–16. https://www.tesolanz.org.nz/news/

Ellis, R. (2010). Second language acquisition, teacher education and language pedagogy. *Language Teaching, 43*(2), 182-201. doi:10.1017/S0261444809990139

Ferris, D. R. (1999). The case for grammar correction in L2 writing classes: A response to Truscott. *Journal of Second Language Writing , 8*(1), 1-10.

Ferris, D. R. (2003). *Response to student writing: implications for second language students*. Lawrence Erlbaum Associates.

Ferris. D. R. (2010). Second language writing research and written corrective feedback in SLA: Intersections and practical applications. *Studies in Second Language Acquisition 32*(2), 181-201. https://doi.org/10.1017/S0272263109990490

Ferris, D. R. & Hedgcock, J. (1998). *Teaching ESL composition : Purpose, process, and practice.* Erlbaum.

Flick, U. (2014). Qualitative content analysis. In U. Flick (Ed.), *The SAGE handbook of qualitative data analysis* (pp.170-183). Sage Publications, Inc.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208-238.

Fulcher, G. (2003). *Testing second language speaking.* Pearson Longman.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29. Doi:10.1177/0265532209359514

Grabe, W. (2004). Research on teaching reading. *Annual Review of Applied Linguistics, 24*, 44-69. DOI: 10.1017/S0267190504000030

Grabe, W. & Kaplan, R. (1996). *Theory and practice of writing: an applied linguistics perspective (Applied Linguistics and Language Study)*, Longman. https://doi.org/10.4324/9781315835853

Green, A. (2014). The Test of English for Academic Purposes (TEAP) Impact Study: Report 1 - Preliminary Questionnaires to Japanese High School Students and Teachers. *Tokyo: Eiken Foundation of Japan.* Retrieved from https://www.eiken.or.jp/teap/group/pdf/teap_washback_study.pdf

Grosz, B.J. & Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics, 12*(3), 175-204. Retrieved from https://aclanthology.org/J86-3001.pdf

Halliday, M.A.K. & Hassan, R. (1976). *Cohesion in English*. Longman.

Halliday, M.A.K. (1985). *An Introduction to Functional Grammar.* Edward Arnold.

Hamp-Lyons, L.(1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second a language writing; Research insights for the classroom* (pp. 69-87). Cambridge University Press. doi:10.1017/CBO9781139524551.009

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts.* Ablex.

Harder, B.D., & Katz-Harder, H. (1982). Cultural interface and teaching English composition in Japan. *The English Teacher's Magazine, 7,* 19-23.

Hatch, E. (1992). *Discourse and language education,* Cambridge Language Teaching Library, Cambridge University Press.

Hawes, T. (2015). Thematic progression in the writing of students and professional, *Ampersand 2,* 93-100. https://doi.org/10.1016/j.amper.2015.06.002

Hinds, J. (1983). Contrastive rhetoric: Japanese and English. *Text, 3*, 183-196. https://doi.org/10.1515/text.1.1983.3.2.183

Hinds, J. (1990). Inductive, deductive, quasi-inductive: Expository writing in Japanese. Korean, Chinese, and Thai. In U. Connor & A.M. Johns (Eds.), *Coherence in writing: Research and pedagogical perspectives* (pp. 87-109). Alexandria, VA: Teachers of English to Speakers of Other Languages (TESOL).

Hirai, A. (2013). *Bunsan-bunseki no Oyo* [Application of Analysis of Variance] for Students, Retrieved 30 Aug 2022, from https://www.u.tsukuba.ac.jp/~hirai.akiyo.ft/forstudents/eigokyouikugaku7.files/2013_6_17.pdf

Hirose, K. (2005). *Product and process in the L1 and L2 writing of Japanese students of English.* Keisuisha.

Hoenishch, S. (1996). The theory and method of topical structure analysis. Retrieved 25 Aug 2020, from http://www.criticism.com/da/tsa-method.php

Horowitz, D. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20* (3)*,* 445-462. https://doi.org/10.2307/3586294

Hyland, K. (2003). *Second Language Writing*. Cambridge University Press.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83-101. https://doi.org/10.1017/S0261444806003399

Im, GH., Shin, D., Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation. *Language Testing in Asia, 9*(14). https://rdcu.be/c3kPG https://doi.org/10.1186/s40468-019-0089-4

Intaraprawat, P. & Steffensen, M. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing, 4*(3), 253-272. https://doi.org/10.1016/1060-3743%2895%2990012-8

Ishii, R. (2006).*Zuhyo no teiji oyobi Dai-ni gengo gakushu-sha no setumei-bun dokkai ni oyobosu eikyo.* [Effects of presentation and completion of charts on second language learners' reading comprehension of expository texts (translation mine)]. *Kyoikugaku Shinri Kenkyu, 54,* 498-508. 石井怜子著。「図表の呈示および完成が第二言語学習者の説明文読解に及ぼす影響—中級後半レベルの成人日本語学習者の場合」教育心理学研究。https://doi.org/10.5926/jjep1953.54.4_498

Jacobs, H., Zinkgraf, S. Wormuth, E., Hartfiel, V., & Hughey, J. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

James, C. (1998). *Errors in language learning and use: Exploring error analysis.* Longman.

Jang, X, & Grabe, W. (2007). Graphic organizers in reading instruction: Research findings and issues. *Reading in a Foreign Language, 19*(1), 34-55.

Jian, X. (2012). Effects of Discourse Structure Graphic Organizers on EFL Reading Comprehension. *Reading in a Foreign Language, 24*(1), 84-105. https://files.eric.ed.gov/fulltext/EJ974105.pdf

Kakai, H. (2015). Kongo kenkyu-ho nyumon: shitsu to ryo ni yoru togo no art [Introduction to mixed methods approach: the art of integration through quality and quantity (translation mine)]. Igaku-shoin. 抱井尚子著。『混合研究法入門—質と量による統合のアート』。医学書院。

Kamimura, T. (1996). Composing in Japanese as a First Language and English as a Foreign Language: a Study of Narrative Writing. *RELC Journal.* https://doi.org/10.1177/003368829602700103

Kane, M. (2006). Validation. In R. Brennan (Ed.). *Educational Measurement* (4th ed., pp. 17-64). American Council on Measurement in Education and Praeger Publishers.

Kaplan, R. (1966). Cultural thought　patterns in intercultural education. *Language Learning, 16,* 1-20. https://doi.org/10.1111/j.1467-1770.1966.tb00804.x

Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second language writing skills, *Modern Language Journal, 75*, 305-313.

Kawanishi, K. (2019). Giron ni ikkansei to kesokusei wo motaseru tameno essay writing shido [Writing instruction to ensure coherence and cohesion in discussions. (translation mine)]. In H. Yamanishi, & J. Otoshi (Eds.), *Chu-Jyo Kyu Eigo Writing Shido Guide* [A Comprehensive Guide to Teaching English Writing Skills to Students] (pp. 136-149). Taishukan.

Kawase, T. (2020). How to Utilize Rhetorical Structure Theory to Teach Paragraph Structure: A Theoretical Review and Proposals for Applications. *JACET Kansai Journal, 22*, 34-54. Retrieved from http://www.jacet-kansai.org/file/2020articles.pdf

Kim, Y. H. (2011). Diagnosing EAP writing ability using the Deduced Reparameterized Unified Model. *Language Testing, 28(4)*, 509-541. DOI: 10.1177/0265532211400860

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Knoch, U. (2007a).'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence, *Assessing Writing 12*(2), 108-128. https://doi.org/10.1016/J.ASW.2007.07.002

Knoch, U. (2007b). Diagnostic writing assessment: the development and validation of rating scale. [Doctoral dissertation, University of Auckland]. Available in ResearchGate https://www.researchgate.net/publication/37986429

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing* 16(2), 81-96. https://doi.org/10.1016/j.asw.2011.02.003

Kobayashi, H. & Rinnert, C. (2002). High school student perceptions of first language literacy instruction: implications for second language writing, *Journal of Second Language Writing, 11*(2), 91-2. https://doi.org/10.1016/S10160-3743(02)0067-X

Kobayashi, H. & Rinnert, C. (2008). Task response and text construction across L1 and L2 writing, *Journal of Second Language Writing, 17*(1), 7-29. http://dx.doi.org/10.1016/j.jslw.2007.08.004

Kools, M. , van de Wiel, M. W., Ruiter, R.A., Cruts, A., & Kok, G. (2006). The effect of graphic organizers on subjective and objective comprehension of a health education text. *Health Education & Behavior,33,* 760-772. doi: 10.1177/1090198106288950. Epub 2006 Sep 26. PMID: 17003246.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18,* 219-240.

Kubota, R. (1998). An investigation of L1-L2 transfer in writing among Japanese university students: Implications for contrastive rhetoric. *Journal of second Language writing, 7*(1), 69-100. Retrieved from https://www.academia.edu/10166804/An_investigation_of_L1_L2_transfer_in_writing_among_Japanese_university_students_Implications_for_contrastive_rhetoric

Kunnan, A. J. & Jang, E.E. (2009). Diagnostic Feedback in Language Assessment, In Long, M.H. & Doughty, C.J. (Eds.), *The Handbook of Language Teaching. Wiley Online Library.* https://doi.org/10.1002/9781444315783.ch32

Kuno, S. (1980). Functional syntax, In E. A. Moravcsik & J. R. Wirth (Eds.), S*yntax and Semantics 13: Current approaches to syntax.* Academic Press.

Larson-Hall, J. (2010) *A Guide to Doing Statistics in Second Language Research Using SPSS. S*econd Language Acquisition Research series. New York. Routledge.

Lautamatii, L. (1987). Observations on the development of the topic of simplified discourse. In U. Connor & R.B. Kaplan (Eds.), *Writing across languages: Analysis of L2 text reading* (pp. 87-114). Addison-Wesley.

Lee, H. (2017). The Effects of University English Writing Classes Focusing on Self and Peer Review on Learner Autonomy. *The Journal of Asia TEEL, 14*(3), 1-18. Retrieved from http://dx.doi.org/10.18823/asiatefl.2017.14.3.6.464

Lee, I. (2002). Teaching coherence to ESL students: A classroom inquiry. *Journal of Second Language Writing, 11*, 135–159. https://doi.org/10.1016/S1060-3743(02)00065-6

Lee, Y. W., & Sawaki, Y. (2009). Cognitive Diagnosis Approaches to Language Assessment: An Overview. *Language Assessment Quarterly, 6*(3), 172-189.

doi: 10.1080/15434300902985108

Leighton, J. P. (2009). Mistaken impressions of large-scale cognitive diagnostic testing. In R. P. Phelps (Ed.), *Correcting Fallacies about Educational and Psychological Testing*. Washington, DC: American Psychological Association, 219–46.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* MESA Press.

Linacre, J. M. (1993). Generalizability Theory and Many-Facet Rasch Measurement. Paper presented at the 1993 Annual Meeting of the American Educational Research Association　Atlanta, Georgia, Retrieved from http://files.eric.ed.gov/fulltext/ED364573.pdf

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2).103-22.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement, 5*(1), 95–110. Retrieved from https://www.winsteps.com/a/Linacre-estimation-further-topics.pdf

Linacre, J.M. (2014). *A user's guide to Winsteps Ministep Rasch-model computer programs (3.81.0).* Retrieved from http://www.womsteps.com/winman/reliability.htm

Linacre, J.M. & Wright, T.F. (1993). *A user's guide to FACETS: Rasch-measurement computer program Version 2.62*. MESA Press.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3). 246-276.

Lyster, R. & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classroom. *Studies in Second Language Acquisition, 19*(1), 37-66.

Maimon, E. P. & Nodine, B. (1978). Measuring syntactic growth: errors and expectations in sentence-combining practice with college freshmen. *Research in the Teaching of English, 12*(3), 233-244.

Mann. W. (2003). RST definitions. Retrieved from http://www.sfu.ca/rst/index.html

Mann, W., Matthiessen, C. & Thompson, S. (1992). Rhetorical Structure Theory and Text Analysis. In W. Mann & S. Thompson (Eds.), *Discourse Descriptions Diverse Linguistic Analyses of a Fund-raising Tex*t, (pp. 39-79). John Benjamins Publishing Company.

Mann, W., & Thompson, S. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization, *Text ---Interdisciplinary Journal for the Study of Discourse, 8*(3), 243-281. Retrieved from https://www.sfu.ca/rst/05bibliographies/bibs/Mann_Thompson_1988.pdf

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Masters, G.N. (2010). The partial credit model. In M.L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 109-122). Routledge.

Matsuda, P. K. (1997). Contrastive Rhetoric in Rhetoric in Context: A Dynamic Model of L2 Writing. *Journal of Second Language Writing, 6*(1), 45-60. https://doi.org/10.1016/S1060-3743(97)90005-9

Matsumura, K. (2020). Bridging EAP (English for Academic Purposes) and EMI (English Medium Instruction)-Needs Analysis Reflecting Students' Perspectives. *Waseda Review of Education, 34*(1), 55-72. Institute for Advanced Studies in Education, Waseda University.

Matsumura, K. & Sakamoto, K. (2021). A Structure Analysis of Japanese EFL Students' Argumentative Paragraph Writings with a Tool for Annotating

Discourse Relations. *The Bulletin of the Writing Research Group, JACET Kansai Chapter, 14,* 31-50.

Matsumura, K. & Takagi, A. (2022). A Qualitative Analysis of Descriptors for the Organization Rating Scales in EFL Essay Writing. *Aoyama Journal of Academic Writing Research, 1*, 45-54. https://www.agulin.aoyama.ac.jp/writingcenter/wp-content/uploads/sites/9/2022/06/598f7c327b29d4f8ab91c1879088999b.pdf

Mauranen, A. (1993). *Cultural Differences in Academic Rhetoric: A Textlinguistic Study*. Peter Lang.

McNamara, T. (1996). *Measuring second language performance*, Longman.

McNamara, T. (2006). Validity in language testing: The Challenge of Sam Messick's Legacy. *Language Assessment Quarterly, 3*(1), 31-51. https://doi.org/10.1207/s15434311laq0301_3

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (pp. 13–103). Macmillan Publishing Co, Inc; American Council on Education.

Messick, S. (1990). Validity of Test Interpretation and Use. Research Report. Educational Testing Service. Princeton, New Jersey, August 1990 https://files.eric.ed.gov/fulltext/ED395031.pdf

Meyers, L. S., Gamst, G., & Guarino, A. J. (2006). *Applied Multivariate Research: Design and Interpretation*. Sage.

Ministry of Education, Culture, Sports, Science and Technology (2018). Heisei 29・30 nen kaitei gakushu shidou youryou: ikiruchikara [National Curriculum Standards (2017-2018 Revision), March, 2018]. Retrieved September 4, 2020, from https://www.mext.go.jp/a_menu/shotou/new-cs/1384661.htm; https://www.mext.go.jp/content/1384661_6_1_3.pdf https://www.mext.go.jp/content/1407073_09_1_2.pdf

Ministry of Education, Culture, Sports, Science and Technology (2020). Basic School Survey. Retrieved September 4, 2020, from https://www.mext.go.jp/content/20200825-mxt_chousa01-1419591_8.pdf

Ministry of Education, Culture, Sports, Science and Technology (2021). [Fostering Comprehensive English Proficiency Reference Material 4, The 21st Study Council on University Entrance Examinations (R3.2.17): Cultivating and Evaluating the Ideal University Entrance Examinations Background (translation mine)]. Retrieved September 4, 2020, from https://www.mext.go.jp/content/20210216-mxt_daigakuc02-000012828_11.pdf

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81–97. https://doi.org/10.1037/h0043158

Miller, D. R. & Bustamante, C. (2016). Drawing mixed methods procedural diagrams, In Moeller, A., Creswell, J. W. & Saville, N. (Eds.), *Second Language Assessment and Mixed Methods Research: Studies in Language Testing, Series* Number 43 (1st ed.)(pp. 84-118). Cambridge University Press.

Mislevy, R. J. (1996). Test Theory Reconceived. *Journal of Educational Measurement , 33*(4).379-416. https://doi.org/10.1111/j.1745-3984.1996.tb00498.x

Moeller, A. J. (2016). The confluence of language assessment and mixed methods, In Moeller, A., Creswell, J. W. & Saville, N. (Eds.), *Second Language Assessment and Mixed Methods Research: Studies in Language Testing, Series* Number 43 (1st ed.), (pp. 3-16). Cambridge University Press.

Montgomery, J. & Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing, 16* (2), 82-99. https://doi.org/10.1016/j.jslw.2007.04.002

Myford, C.M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education, 15*(2), 187-215.

Nishigaki, C., Chujo, K., McGoldrick, S., & Hasegawa, S. (2007). A Cross-Sectional Contrastive Analysis of Japanese Students' English Composition Skills. *THE JOURNAL OF ASIA TEFL, 4*(1), 27-54. Retrieved from http://www5d.biglobe.ne.jp/~chujo/data/Chikako_Nishigaki.pdf

Norris, J. (2009). Task-Based Teaching and Testing. In Long, M. & Doughty, C. (Eds.), *The Handbook of Language Teaching* (pp.578-594). https://doi.org/10.1002/9781444315783.ch30

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23*(4), 445-465.

North, B. (2000). *The development of a common framework scale of language proficiency.* Peter Lang.

North, B. (2003). Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. *TOEFL Monograph 24.* Educational Testing Service. Retrieved from https://www.researchgate.net/publication/291770049_Scales_for_rating_language_performance_Descriptive_models_formulation_styles_and_presentation_formats#fullTextFileContent

North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217-263.

Luoma S (Eds.), Current developments and alternatives in language assessment. *Proceedings of the LTRC 1996* (pp. 423–447). Jyvaskyla: University of Jyvaskyla Press.

O'Brien, T. (1995). Rhetorical structure analysis and the case of the inaccurate, incoherent source-hopper. *Applied Linguistics, 16*(4), 442-482.

doi: 10.1093/applin/16.4.442

Oi, K. (1986). Cross-cultural Differences in Rhetorical Patterning--A Study of Japanese and English. *JACET Bulletin*, *17*, 23-48. Retrieved from https://dl.ndl.go.jp/pid/10500387/1/1

Oi, K. (1999). Comparison of argumentative styles: Japanese college students vs. American college students--An analysis using the Toulmin model. *JACET Bulletin, 30*, 85-102. Retrieved from https://dl.ndl.go.jp/pid/10501296/1/1

Oi, K. (2005). Teaching argumentative writing to Japanese EFL students using the Toulmin model. *JACET Bulletin, 41*, 123-140. Retrieved from https://dl.ndl.go.jp/pid/10501483/1/1

Oi, K. & Horne, B. (2016). Comparison of EFL Writing Practices Perceived through a Students' Survey between Japan, South Korea, Taiwan and Hong Kong. In K. Oi (Ed.), *EFL Writing in East Asia: Practice, Perception and Perspectives* (pp.64-94). Shobi Printing.

Oi, K., Itatsu, Y. & Horne, B. (2016). A Survey of Awareness of and Attitudes to EFL Writing among Junior and Senior High School Teachers: Perspectives Gained through a Comparison between Japan, Korea and Taiwan. In K. Oi (Ed.), *EFL Writing in East Asia: Practice, Perception and Perspectives* (pp.14-38). Shobi Printing.

Oi, K., Kamimura, T., Kumamoto, T., & Matsumoto, K. (2000). A Search for the feedback that works for Japanese EFL students: Content-based or grammar-based. *JACET Bulletin, 32*,91-108. https://dl.ndl.go.jp/pid/10501330/1/1

Oi, Y. (2021). Efficacy of Student Assessment as Part of English Writing Instruction for Japanese High School Students. [Unpublished Doctoral dissertation]. Waseda University. https://waseda.repo.nii.ac.jp/?action=pages_view_main&active_action=reposito

ry_view_main_item_detail&item_id=65733&item_no=1&page_id=13&block_id=21

Olson, C.L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin, 83,* 579-586. doi: 10.1111/lang.12079

Pugh, D., Hickson, D., Hinings, C. & Turner, C. (1969) The Context of Organization. *Administrative Science Quarterly 14* (1), 91-114. https://doi.org/10.2307/2391366

Putra, J. W. G., Teufel, S., and Tokunaga, T. (2019). An Argument Annotation scheme for the Repair of Student Essays by Sentence Reordering. In *Proceedings of Annual Meeting of the Association for Natural Language Processing Japan*, pp. 546–549. https://www.anlp.jp/proceedings/annual_meeting/2019/pdf_dir/P3-9.pdf

Putra, J. W.G., Teufel, S., Matsumura, K. & Tokunaga, T. (2020). TIARA: A Tool for Annotating Discourse Relations and Sentence Reordering. *Proceedings of the 12rh International Conference on Language Resources and Evaluation (LREC)*, pp. 6912–6920. https://aclanthology.org/2020.lrec-1.854.pdf

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4-13.https://www.researchgate.net/publication/227634769_On_the_Definition_of_Feedback http://dx.doi.org/10.1002/bs.3830280103

Richards, J. C. (1990). *The Language Teaching Matrix.* New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511667152

Robinson, D.H. & Kiewra, K.A. (1995). Visual argument: Graphic organizers are superior to outlines in improving learning from text. *Journal of Educational Psychology, 87*(3), 455-467. https://psycnet.apa.org/doi/10.1037/0022-0663.87.3.455

Rosenfeld, M., Leung, S., & Oltman, P.K. (2001). *The Reading, Writing, Speaking, and Listening Tasks Important for Academic Success at the Undergraduate and Graduate Levels.* (TOEFL report no. MS-21). Princeton: ETS.

Sadler, D.R. (1989). Formative assessment and the design of instructional systems. *Instructional Science 18,* 119–144. https://doi.org/10.1007/BF00117714

Sakamoto, K. (2016). *Applying Toulmin's Argument Model to English Writing Instruction*. [Unpublished master's thesis], Kyoto University.

Sasaki, M. & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning, 46(*1), 137-168. https://doi.org/10.1111/j.1467-1770.1996.tb00643.x

Saville, N. (2016). Managing language assessment systems and mixed methods, In Moeller, A., Creswell, J. W. & Saville, N. (Eds.), *Second Language Assessment and Mixed Methods Research: Studies in Language Testing, Series Number 43* (1st ed.)(pp. 17-31). Cambridge University Press.

Sawaki, Y. (2011). Daikibo gengo test no datosei yuyosei ni kansuru kin-nen no doko [Recent Trends in Examining the Validity and Usefulness of Large-Scale Linguistic Tests (translation mine)]. *Assessment and Evaluation in Language Education, 2*, 54-63. Retrieved from https://obirin.repo.nii.ac.jp/?action=repository_action_common_download&item_id=1255&item_no=1&attribute_id=21&file_no=1

Sawaki, Y. (2017). University faculty members' perspectives on English language demands in content courses and a reform of university entrance examinations in Japan: a needs analysis. *Language Testing in Asia, 7*: 13. https://doi.org/10.1186/s40468-017-0043-2

Schneider, M., & Connor, U. (1990). Analyzing topical structure in ESL essays. *Studies in Second Language Acquisition 12*(4*),* 411-427.

Shank, R. & Abelson, R. (1977). *Scripts, plans, goals and understanding: an inquiry into human knowledge structures.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sheppard, K. (1992). Two feedback types: Do they make a difference? *RELC Journal, 23*, 103-110. https://doi.org/10.1177/003368829202300107

Shimabayashi, S. (2019). Ronsho-bun wo kaku [Writing argumentative paragraphs. (translation mine)]. In H. Yamanishi, & J. Otoshi (Eds.), *Chu-Jyo Kyuu Eigo Raiting Shido Gaido* [A Comprehensive Guide to Teaching English Writing Skills to Students] (pp. 122-135). Taishukan.

Shohamy, E. & Hornberger, N.H. (2008). *Language Testing and Assessment: Encyclopedia of Language and Education Volume 7.* Springer Science Business Media LLC.

Simpson, J. M. (2000). Topical structure analysis of academic paragraphs in English and Spanish. *Journal of Second Language Writing, 9*(3)*,* 293-303.

Skoufaki, S. (2009). An exploratory application of rhetorical structure theory to detet coherence errors in L2 English writing: Possible implications for automated writing evaluation software. *International Journal of Computational Linguistics and Chinese Language Processing: Special Issue in Computer Assisted Language Leaning, 14*, 181-203. Retrieved from https://aclanthology.org/O09-4003

Smith, E.V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*(2)*,* 205-231.

Sonntag, J. & Stede, M. (2014). GraPAT: a tool for graph annotations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 4147 – 4151.

Spillner, B. (1991). *Error analysis: a comprehensive bibliography.* John Benjamins Pub. Co.

Stevens, J.P. (1980). Power of the multivariate analysis of variance test, *Psychological Bulletin, 88*, 728-737. doi: 10.1037/0033-2909.88.3.728

Storch, N. (2010). Critical feedback on written corrective feedback research. *International Journal of English Studies*, *10*(2), 29-46. https://doi.org/10.6018/ijes/2010/2/119181

Suzuki, T., Oi, K. & Takemae, F. (Eds.). (2006). *Critical thinking to kyoiku: Nihon no kyoiku wo saikochiku suru [Critical thinking and education : Reconstructing Japanese education]*. Sekai Shiso-sha.

Taboada, M. & Mann. W.C. (2006). Applications of Rhetorical Structure Theory. *Discourse Studies, 8*(4), 567-588. http://dx.doi.org/10.1177/1461445606064836

Takagi, A. (2021, July 16). [PowerPoint slides]. *Qualitative analysis of results - thematic analysis* presented in Workshop hosted by JACET SIG on English Language Education. http://www.waseda.jp/assoc-jacetenedu/

Takahara, P.O. (2003). Micropragmatics--grammar and pragmatics. In I. Koike, S. Ide, M. Kono, H. Suzuki, H. Tanaka, S. Tanabe., & O. Mizutani (Eds.) *Kenkyusha Dictionary of Applied Linguistics* (pp. 278-279). Tokyo: Kenkyusha.

Takanami, S. (2012). [Chapter 6:Application of Analysis of Variance: Covariance Analysis and Multivariate Analysis of Variance (pp.121-144). In Hirai, A. (ed.), *Introduction to Data Analysis for Educational and Psychological Research: SPSS Application Methods in Theory and Practice.* Tokyo Shoseki. (translation mine)]

Tanaka, M. (2015). Writing kenkyu to feedback. In Ozeki (Ed.) *Feedback Kenkyu he no Shotai* [An Invitation to Feedback Research: Understanding Feedback in SLA] (pp. 107-138). Kuroshio shuppan.

Todd, R.W., Thienpermpool, P. & Keyuravong, S. (2004). Measuring the coherence of writing using topic-based analysis, *Assessing Writing, 9*(2), 85-104. https://doi.org/10.1016/j.asw.2004.06.002

Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.

Tomita, F. (2019). *Eiken writing mondai no shido 2 kyu, jun 2kyu, 3 kyu no gaiyo to taisaku* [Teaching EIKEN Writing Test Topics-Grade 2, Pre 2, and Grade 3-Overview and Strategies (translation mine)] . In Yamanishi, H. & Otoshi, J. (Eds.). *A Comprehensive Guide to Teaching English Writing Skills to Students* (pp.258-285). Taishukan shoten.

Truscott, J. (1996). The case against grammar correction in L2 writing classes. *Language Learning, 46*(2), 327-369. https://doi.org/10.1111/j.1467-1770.1996.tb01238.x

Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. J*ournal of second Language Writing, 16*(4), 255-272. https://doi.org/10.1016/j.jslw.2007.06.003

Tsuji, K. (2016). Teaching argumentative writing through a process-focused instruction: The effects of the prewriting activity on student perceived learning. *Kyoto University's Library of Higher Education Research, 22,* 77-86. https://repository.kulib.kyoto-u.ac.jp/dspace/handle/2433/219549

Turner, C.E., & Upshur, J.A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student samples on scale content and student scores. *TESOL Quarterly, 36*, 49-70. Doi:10.2307/3588360

Upshur, J.A., & Turner, C.E. (1996). Constructing rating scales for second language tests. *ELT Journal, 59*, 3-12. Doi:10.1093/let/49.1.3

Vande Kopple, W. J. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication, 36,* 82-93. https://eric.ed.gov/?id=EJ311449

van Dijk, T.A. (1977). *Text and Context: Explorations in the Semantics and Pragmatics of Discourse.* Addison-Wesley Longman.

van Dijk, T.A. (1980). *Macrostructures. An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition.* Lawrence Erlbaum Associates.

Vekiri, I. (2002). What Is the Value of Graphical Displays in Learning? *Educational Psychology Review, 14*(3), 261-312. https://doi.org/10.1023/A:1016064429161

Wagner, M. (2015). The Centrality of Cognitively Diagnostic Assessment for Advancing Secondary School ESL Students' writing: A Mixed Methods Study. [Doctoral dissertaion, University of Toronto]. Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/69530/3/Wagner_Maryam_201506_PhD_thesis.pdf

Wang, L. (2007). Theme and Rheme in the thematic organization of text: Implications for teaching academic writing. *Asian TEFL Journal, 9*, 164-176.

Watson Todd, R., Khongput, & Darasawang (2007). Coherence, cohesion and comments on students' academic essays. *Assessing writing, 12*(1), 10-25. https://doi.org/10.1016/j.asw.2007.02.002

Weir, C.J. (1988). Construct validity, In A. Hughes, D. Porter & C. J. Weir (Eds.), *ELTS Validation project report (ELTS Research reports I (ii))*. London: The British Council/UCLES.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach.* Palgrave Macmillan.

White, E.M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating and improving student performance.* (2nd ed.). San Francisco: Jossey-Bass.

Wigglesworth, G. (2008). Task and performance-based assessment, In Shohamy, E. & Hornberger, N.H. (Eds.). *Encyclopedia of Language and Education (2nd ed.), Language Testing and Assessment vol. 7* (pp. 11-122). Springer Science Business Media LLC.

Wikborg, E. (1990). Types of coherence breaks in Swedish student writing : Misleading paragraph division . In U. Connor & A.M. Johns (Eds.), *Coherence in writing: Research and pedagogical perspectives* (pp. 131-148). TESOL.

Wilkinson, A. M. (1991). *The scientist's handbook for writing papers and dissertations.* Edgewood Cliffs, NJ: Prentice Hall.

Winn, W., Li, TZ., & Schill, D. (1991). Diagrams as aids to problem solving: Their role in facilitating search and computation. *Educational Technology Research and Development, 39*(1), 17-29. https://doi.org/10.1007/BF02298104

Witte, S. (1983a). Topical Structure and Revision: An Exploratory Study. *College Composition and Communication, 34(3)*, 313-341.

Witte, S. (1983b). Topical Structure and Writing Quality: Some Possible Text-Based Explanation of Readers' Judgments of Students' Writing . *Visible Language, 17*, 177-205.

Witte, S. & Faigley, L. (1981). Coherence, cohesion , and writing quality. C*ollege Composition and Communication, 32,* 189-204.

Wolf, F. & Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics, 31(2),* 249–288.

Yamanishi, H. (2019). *Nihonjin eigo gakushusha ni taisuru writing kenkyu shido jissen no tenkai -What, How, and Why no wakugumi wo enyo shite-* [Writing Research and Instructional Practices for Japanese Learners of English: Using the What, How, and Why Framework (translation mine)]. In Yamanishi, H. & Otoshi, J. (Eds.). *A Comprehensive Guide to Teaching English Writing Skills to Students* (pp. 4-14). Taishukan shoten.

Yamanishi, H. & Tanaka, H. (2003). Combination of Qualitative and Quantitative Researches in English Language Education Study: From Hypothesis-testing to Hypothesis-succeeding. *Language Education & Technology, 40*, 161-173. https://doi.org/10.24539/let.40.0_161

Yamashita, M. (2019). An Analysis of Rhetorical Features and Logical Anomalies in the EFL Argumentative Essays Written by Japanese University Students. [Unpublished Doctoral dissertation]. Kansai University. http://doi.org/10.32286/00018645

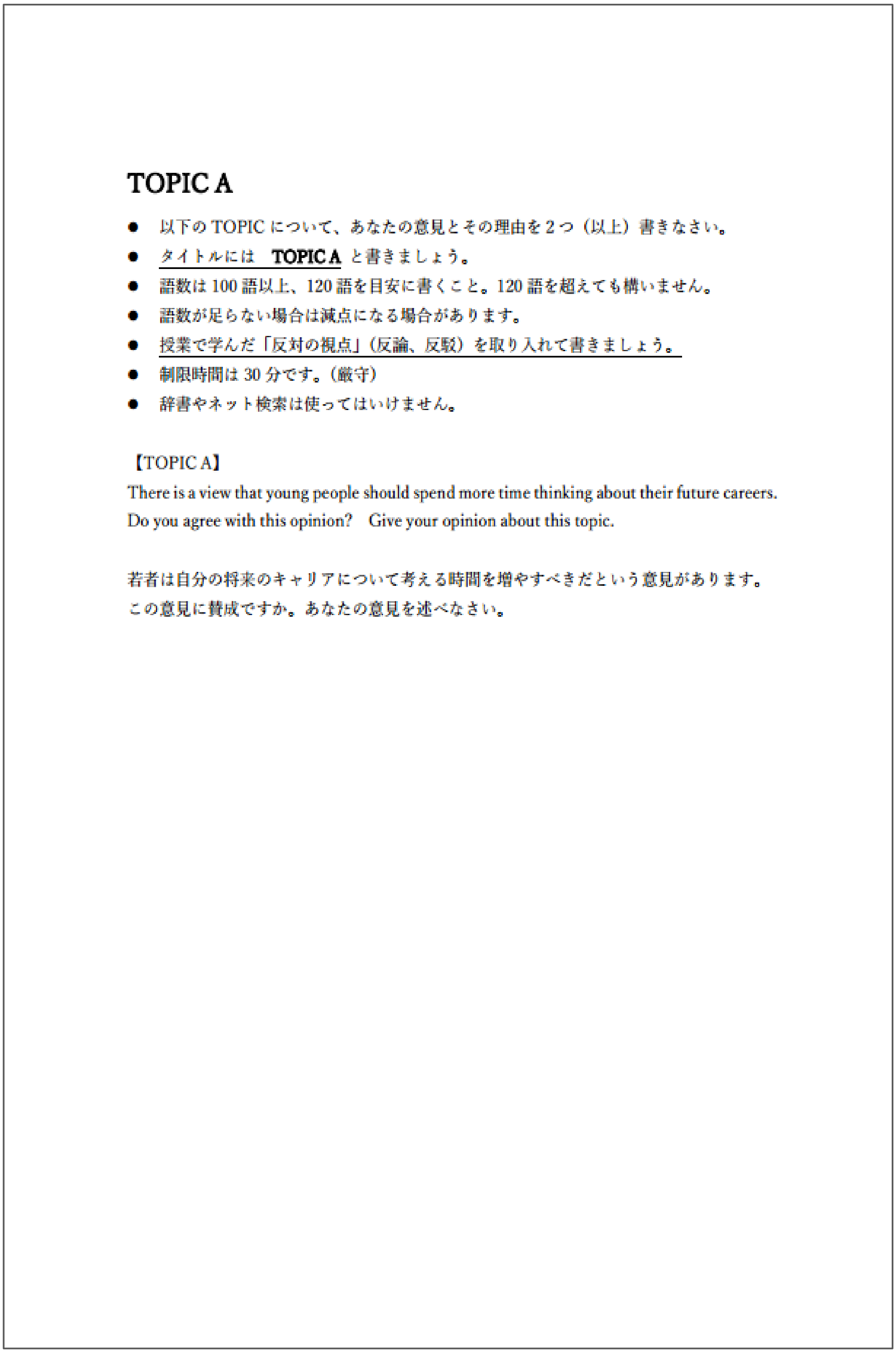
Yasuda, S. (2006). Japanese students' argumentative writing in English: Characteristic weaknesses and developmental factors. *KATE Bulletin, 20*(1), 1-12. DOI:10.20806/katejo.20.0_1

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly19*(1), 79-101. https://doi.org/10.2307/3586773

Zeigler, N. & Kang, L. (2016). Mixed methods designs. In Moeller, A., Creswell, J. W. & Saville, N. (Eds.), *Second Language Assessment and Mixed Methods Research: Studies in Language Testing, Series Number 43* (1st ed.)(pp. 51-83). Cambridge University Press.

Appendix A

A Sample: Instruction for task A (pre)



TOPIC A

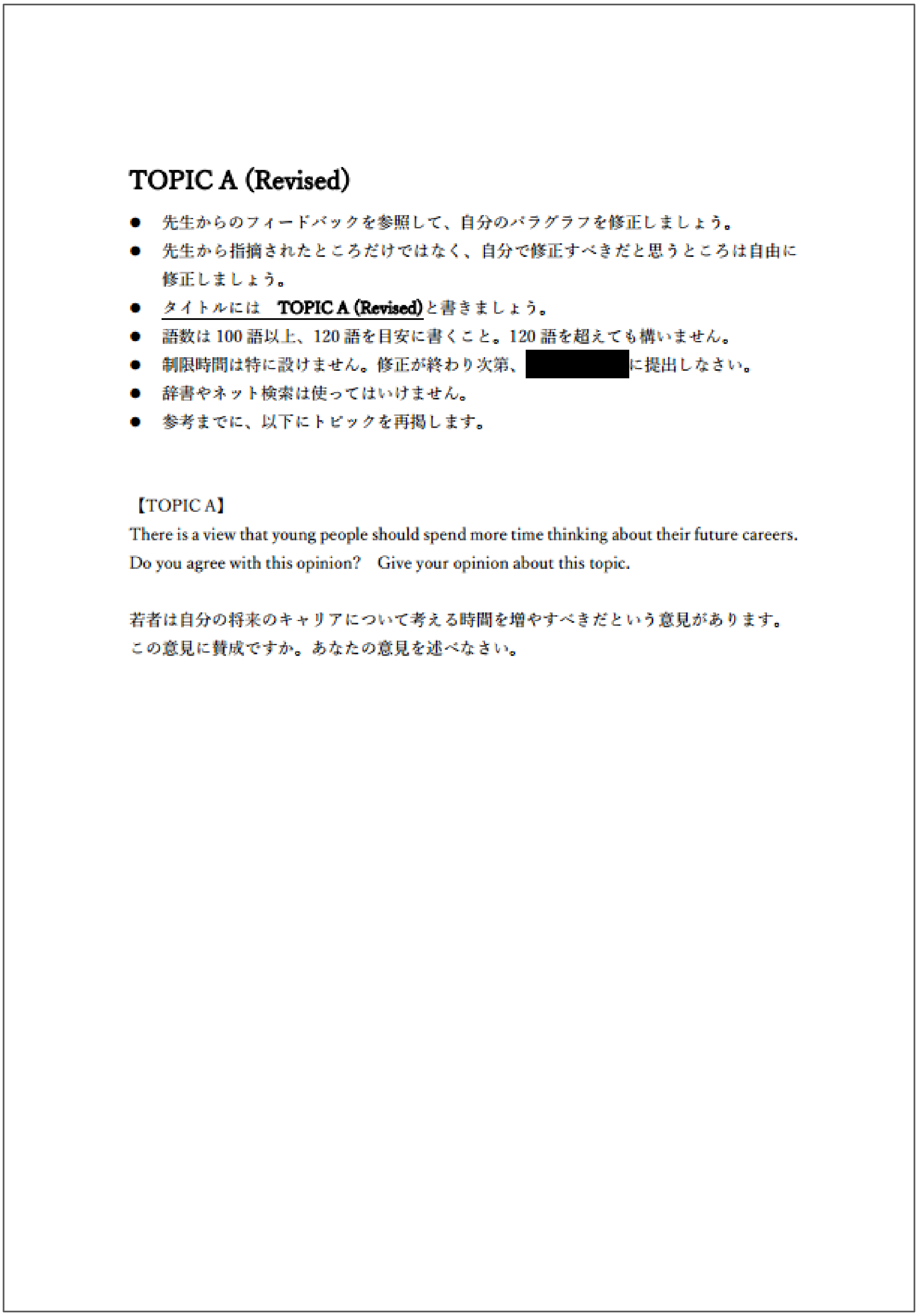
- 以下のTOPICについて、あなたの意見とその理由を２つ（以上）書きなさい。
- タイトルには **TOPIC A** と書きましょう。
- 語数は100語以上、120語を目安に書くこと。120語を超えても構いません。
- 語数が足らない場合は減点になる場合があります。
- 授業で学んだ「反対の視点」（反論、反駁）を取り入れて書きましょう。
- 制限時間は30分です。（厳守）
- 辞書やネット検索は使ってはいけません。

【TOPIC A】
There is a view that young people should spend more time thinking about their future careers.
Do you agree with this opinion? Give your opinion about this topic.

若者は自分の将来のキャリアについて考える時間を増やすべきだという意見があります。
この意見に賛成ですか。あなたの意見を述べなさい。

Appendix B

A Sample: Instruction for task A (post)



TOPIC A (Revised)

- 先生からのフィードバックを参照して、自分のパラグラフを修正しましょう。
- 先生から指摘されたところだけではなく、自分で修正すべきだと思うところは自由に修正しましょう。
- タイトルには　**TOPIC A (Revised)** と書きましょう。
- 語数は 100 語以上、120 語を目安に書くこと。120 語を超えても構いません。
- 制限時間は特に設けません。修正が終わり次第、[redacted]に提出しなさい。
- 辞書やネット検索は使ってはいけません。
- 参考までに、以下にトピックを再掲します。

【TOPIC A】
There is a view that young people should spend more time thinking about their future careers. Do you agree with this opinion?　Give your opinion about this topic.

若者は自分の将来のキャリアについて考える時間を増やすべきだという意見があります。この意見に賛成ですか。あなたの意見を述べなさい。

Appendix C

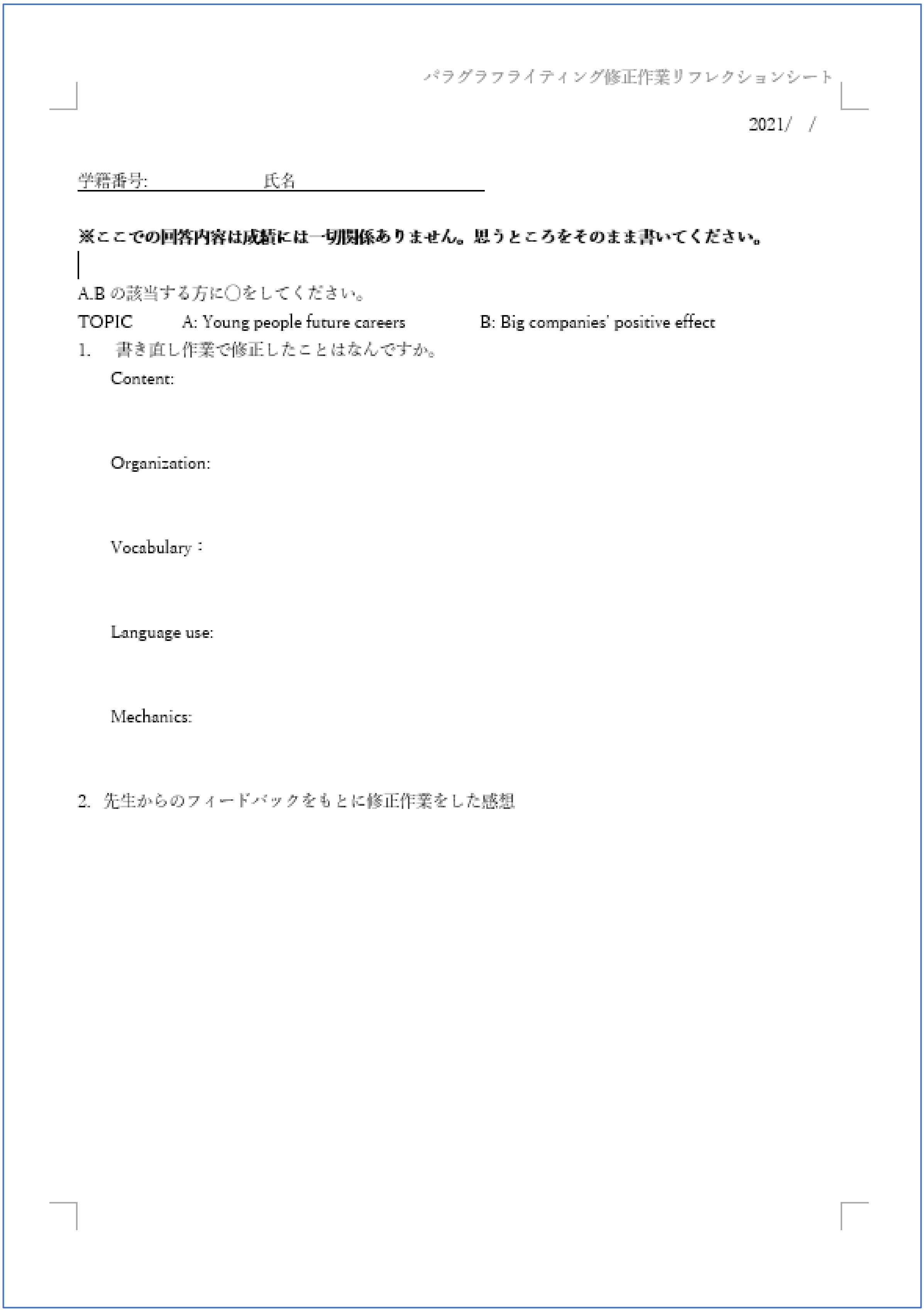
Scoring Sheet for Raters (Simplified Version)

NUMBER:			DATE:
	Score	LEVEL	CRITERIA
CONTENT		25-22	**EXCELLENT TO VERY GOOD**: 知識がある・しっかりした根拠がある・トピックセンテンスが十分に展開している・課題トピックと合致している
		21-18	**GOOD TO AVERAGE:** 主題についていくらかの知識がある・まずまずの広がりがある・トピックセンテンスの展開が限定的である ・大半がトピックに関連しているが、詳細な説明に欠ける・根拠が2つ以上示されていない
		17-11	**FAIR TO POOR**: 主題についての知識が限られている・根拠がほとんどない・トピックの展開が不十分である・
		10-5	**VERY POOR**: 主題についての知識が示されていない・根拠がない・関連がない・評価するには分量が足りない
ORGANIZATION		25-22	**EXCELLENT TO VERY GOOD**: 表現がよどみなく流れている・主張が明確に述べられ裏付けられている・簡潔である・統一性がある・論理的な順序・結束性がある・反論反駁の展開の成功
		21-18	**GOOD TO AVERAGE**: 幾分つながりが悪い・統一性が緩いが、伝えたいideasははっきりわかる ・裏付けが一部に限定されている・論理的だが順序には不完全な点がある ・繰り返しや冗長な文が含まれている場合がある
		17-11	**FAIR TO POOR**: 流れがあちこちで断ち切られる・ideaが混乱しているか、つながっていない・論理的順序と展開に欠ける・突飛あるいは意味が不明な文が流れを阻害している
		10-5	**VERY POOR**: 伝えたいことがわからない・ 統一性が全くない・ 評価するには分量が足りない
	COMMENTS on Organizaiton		
LANGUAGE		25-22	**EXCELLENT TO VERY GOOD**: 効果的で複合的な構造になっている・一致、時制、数、語の順序／機能、冠詞、代名詞、前置詞について、誤りがほとんどない・タイトルがキャピタライズ、中央揃えで表記される
		21-18	**GOOD TO AVERAGE**: 効果的だが単純な構造である・複合的な構造における小さい問題がある・一致、時制、数、語の順序／機能、冠詞、代名詞、前置詞について、誤りがいくつかあるが、意味は十分に伝わる
		17-11	**FAIR TO POOR**: 単純な／複合的な構造における大きな問題がある ・否定、一致、時制、数、語の順序／機能、冠詞、代名詞、前置詞について誤りがしばしば見られる・不完全文、無終止文、欠落がある・意味がわからない、またははっきりしない
		10-5	**VERY POOR**: 文の構造についてのルールがほとんどわかっていない・ 誤りがほとんどである ・ 意味が伝わらない・ 評価するには分量が足りない
VOCABULARY		20-18	**EXCELLENT TO VERY GOOD**: 語彙が洗練された域に達している・単語と成句が効果的に選択され使用されている・語形について熟達している・適切な言語使用域を用いている・日本語、略語の扱いが適切である
		17-14	**GOOD TO AVERAGE**: 語彙の多さが十分である・単語と成句の形、選択、使用について時に誤りがあるが、意味は十分に伝わる
		13-10	**FAIR TO POOR**: 語彙が限られている・単語と成句の形、選択、使用についてしばしば誤りがある・意味がわからない、またははっきりしない
		9-7	**VERY POOR**: 母語の置き換えにすぎない・ 英語の語彙、成句、語形の知識がほとんどない・ 評価するには分量が足りない
MECHANICS		5	**EXCELLENT TO VERY GOOD**: 慣行に習熟していることを示している ・スペル、句読法、大文字使用、段落分けについて、誤りがほとんどない・適切な改行、字下げ・タイトルのキャピタライズ、中央揃え
		4	**GOOD TO AVERAGE**: スペル、句読法、大文字使用、段落分けについて、時に誤りがあるが、意味は十分に伝わる
		3	**FAIR TO POOR**: スペル、句読法、大文字使用、段落分けについて、誤りがしばしば見られる ・
		2	**VERY POOR**: 慣行を全く身につけていない・ スペル、句読法、大文字使用、段落分けについて、誤りがほとんどである
TOTAL	0	**READER (name)**	
		Comments	

Note. Based on Jacobs et al. (1981). Translated by the author.

Appendix D

Questionnaire Form



パラグラフライティング修正作業リフレクションシート

2021/　/

学籍番号:　　　　　　　氏名

※ここでの回答内容は成績には一切関係ありません。思うところをそのまま書いてください。

A.B の該当する方に〇をしてください。

TOPIC　　A: Young people future careers　　　B: Big companies' positive effect

1. 書き直し作業で修正したことはなんですか。

   Content:

   Organization:

   Vocabulary：

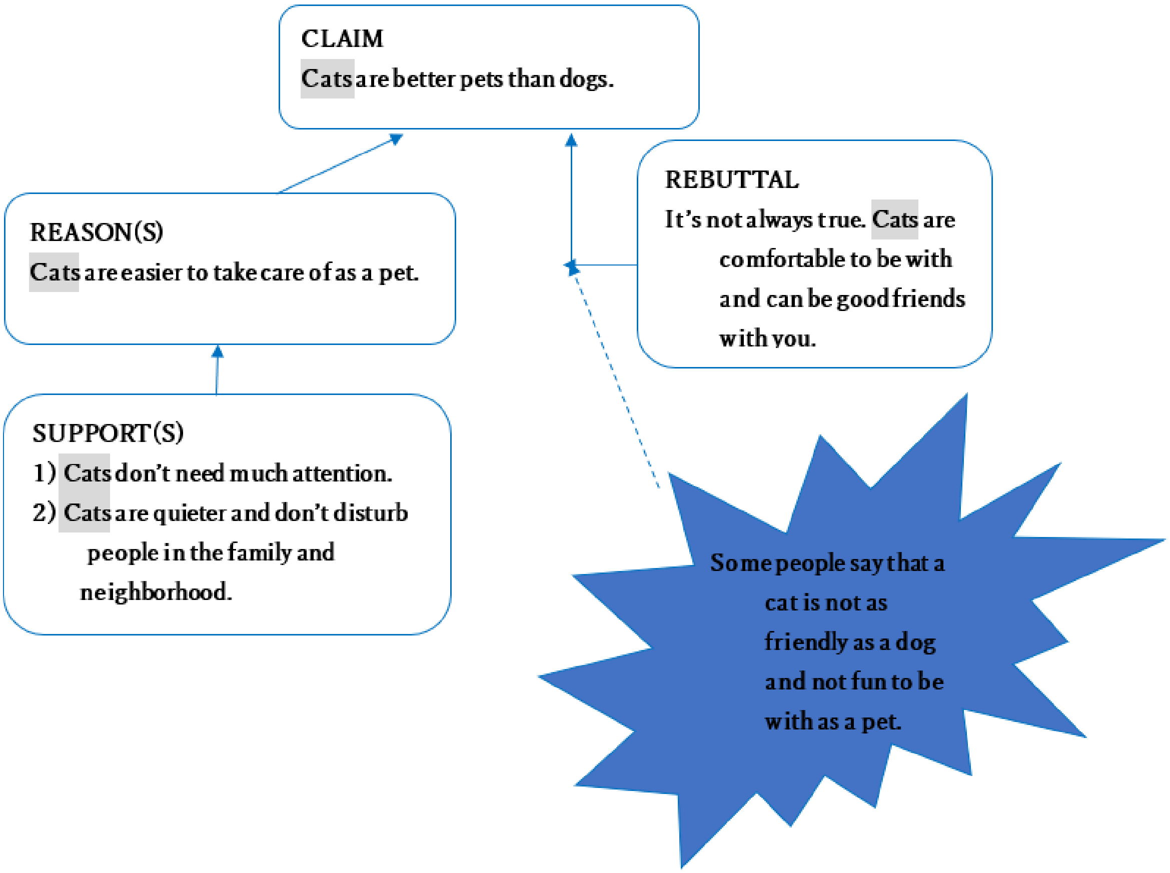
   Language use:

   Mechanics:

2. 先生からのフィードバックをもとに修正作業をした感想

Appendix E

A material for Instruction of Argumentation: Conceptual illustration



CLAIM
Cats are better pets than dogs.

REASON(S)
Cats are easier to take care of as a pet.

SUPPORT(S)
1) Cats don't need much attention.
2) Cats are quieter and don't disturb people in the family and neighborhood.

REBUTTAL
It's not always true. Cats are comfortable to be with and can be good friends with you.

Some people say that a cat is not as friendly as a dog and not fun to be with as a pet.

Appendix F

A Material for Instruction of Argumentation : Text Comparison

A ：Watching the movie with subtitles is better than dubbed ones. Watching movies in subtitles has several advantages. You can listen to actors’ voices as they are. It helps us feel the atmosphere of the movie. Voice is an important part of acting. Moreover, you can learn the language by comparing the utterance and the subtitles. Therefore, I think subtitle versions are better than dubbed versions.

映画は字幕で観る方が吹き替えより良い。字幕だと、役者の声がそのまま聴けるし、外国語の勉強にもなる。役者の声がそのまま聴けると映画の雰囲気が損なわれない。役者の声も含めて大切な演技だと思う。それに、台詞と字幕を比べて語学の勉強にもなる。だから字幕版の方が吹き替え版より良いと思う。

B ：Watching the movie in dubbed version is better than subtitle ones. I believe you can appreciate the movie much better. You can listen in your native language, so you perfectly understand every word they utter. In dubbed version, you don’ have to read the subtitles so you never miss the action That makes you pay attention to the story. Some people say you can’t enjoy the actor’s voices, but voice actors are trained and such a specialist that they will not be distracting. So, dubbed versions are better.

吹き替えで映画を観る方が字幕より良い。映画をもっと楽しめると思う。自分の母語で聴けるので一言一句全部理解できる。吹き替えだと字幕を読まなくて良いので動きを見逃すとがなく、ストーリーにも集中できる。実際の役者の声を楽しめないという人もいるが、声優は訓練を受けた専門家なので気が散ることもない。なので、吹き替えの方が良い。